Analyse de l'illumination et des propriétés de réflectance en utilisant des collections d'images

Mauricio Diaz

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministérial : 7 août 2006

Présentée par

Mauricio Díaz

Thèse dirigée par **Peter Sturm**

préparée au sein **du Laboratoire Jean Kuntzmann (LJK), l'INRIA Grenoble Rhône-Alpes**
et de **Mathématiques, Sciences et Technologies de l'Information, Informatique**

Analyse de l'Illumination et des Propriétés de Reflectance en Utilisant des Collections d'Images

Thèse soutenue publiquement le **26 Octobre 2011**,
devant le jury composé de :

Joëlle Thollot
Professeur, Université de Grenoble, Présidente
Theo Gevers
Professeur, University of Amsterdam, Rapporteur
Ludovic Macaire
Professeur, Université Lille 1, Sciences et Technologies, Rapporteur
Slobodan Ilic
Professeur Adjoint, Technische Universität München, Examinateur
Peter Sturm
Directeur de Recherche, INRIA Grenoble Rhône–Alpes, Directeur de thèse

Analyse de l'Illumination et des Propriétés Photométriques en Utilisant des Collections d'Images

Resumé: L'utilisation de collections d'images pour les applications de vision par ordinateur devient de plus en plus commune des nos jours. L'objectif principal de cette thèse est d'exploiter et d'extraire des informations importantes d'images de scènes d'extérieur à partir de ce type de collections : l'illumination présente au moment de la prise, les propriétés de réflectance des matériaux composant les objets dans la scéne et les propriétés radiométriques des appareils photo utilisés.

Pour atteindre notre objectif, cette thèse et composé de deux parties principales. Dans un premier temps nous allons réaliser une analyse de différentes représentations du ciel et une comparaison des images basée sur l'apparence de celui–ci. Une grande partie de l'information visuelle perçue dans les images d'extérieures est due à l'illumination en provenance du ciel. Ce facteur est représenté par les rayons du soleil réfléchis et réfractés dans l'atmosphère en créant une illumination globale de l'environnement. En même temps cet environnement détermine la façon de percevoir les objets du monde réel. Étant donné l'importance du ciel comme source d'illumination, nous formulons un processus générique en trois temps, segmentation, modélisation et comparaison des pixels du ciel, pour trouver des images similaires en se basant sur leurs apparences. Différentes méthodes sont adoptées dans les phases de modélisation et de comparaison. La performance des algorithmes est validée en trouvant des images similaires dans de grandes collections de photos.

La deuxième partie de cette thèse consiste à exploiter l'information géométrique additionnelle pour en déduire les caractéristiques photométriques de la scène. A partir d'une structure 3D récupérée en utilisant des méthodes disponibles, nous analysons le processus de formation de l'image à partir de modèles simples, puis nous estimons les paramètres qui les régissent. Les collections de photos sont généralement capturées par différents appareils photos, d'où l'importance d'insister sur leur calibrage radiomètrique. Notre formulation estime cet étalonnage pour tous les appareils photos en même temps, en utilisant une connaissance a priori sur l'espace des fonctions de réponse des caméras possibles. Nous proposons ensuite, un cadre d'estimation conjoint pour calculer une représentation de l'illumination globale dans chaque image, l'albedo de la surface qui compose la structure 3D et le calibrage radiométrique pour tous les appareils photos.

Mots Clées: Vision par Ordinateur, Photométrie, Illumination, Calibrage Radiométrique, Collections de Photos.

Estimating Illumination and Photometric Properties using Photo Collections

Abstract: The main objective of this thesis is to exploit the photometric information available in large photo collections of outdoor scenes to infer characteristics of the illumination, the objects and the cameras.

To achieve this goal two problems are addressed. In a preliminary work, we explore optimal representations for the sky and compare images based on its appearance. Much of the information perceived in outdoor scenes is due to the illumination coming from the sky. The solar beams are reflected and refracted in the atmosphere, creating a global illumination ambience. In turn, this environment determines the way that we perceive objects in the real world. Given the importance of the sky as an illumination source, we formulate a generic 3–step process in order to compare images based on its appearance. These three stages are: segmentation, modeling and comparing of the sky pixels. Different approaches are adopted for the modeling and comparing phases. Performance of the algorithms is validated by finding similar images in large photo collections.

A second part of the thesis aims to exploit additional geometric information in order to deduce the photometric characteristics of the scene. From a 3D structure recovered using available multi–view stereo methods, we trace back the image formation process and estimate the models for the components involved on it. Since photo collections are usually acquired with different cameras, our formulation emphasizes the estimation of the radiometric calibration for all the cameras at the same time, using a strong prior on the possible space of camera response functions. Then, in a joint estimation framework, we also propose a robust computation of the global illumination for each image, the surface albedo for the 3D structure and the radiometric calibration for all the cameras.

Keywords: Computer Vision, Photometry, Illumination, Radiometric Calibration, Photo Collections.

Acknowledgments

Il n'y a pas de mots pour exprimer ma gratitude à mon directeur de thèse, Peter Sturm. Il a encadré mes travaux de recherche d'une façon magnifique, en trouvant toujours l'équilibre parfait entre la liberté pour proposer de nouvelles idées et le guidage nécessaire pour réussir dans mes objectives. Merci beaucoup!
I want to thank to the reviewers and the juries, who kindly accepted to review this thesis. I hope you enjoy to read it as I enjoyed these wonderful years as Ph.D student.
Special thanks to the professors Carlos Alberto Parra and Pedro Raul Vizcaya from the Javeriana University at Bogotá, who introduced me to the research world and gave me priceless advices.
I also thanks the Program Alban[1], a funding research program of the European Union who partially fund my thesis work. Particularly, I want to thank to all the administrative staff, who were extremely diligent every time I need them.
To the numerous colleagues and officemates I had over the years at INRIA Grenoble, specially to the members of Perception and Steep, thanks guys!
A mis padres por haberme inculcado desde muy pequeño la curiosidad necesaria para entrar en el mundo de la ciencia, a mi hermana por cuidar y llevar por el buen camino al bejamín y a la tía y la abuelita por todos sus cuidados.
Finalmente, este trabajo de tesis está dedicado a Marie, que aceptó la loca idea de comenzar una aventura de mi lado, y que ha creído siempre en mi. Hemos llegado lejos y estoy seguro que aún nos faltan muchos caminos por recorrer. Sin tu soporte, llegar al final de esta etapa hubiera sido muy difícil. Gracias! Y evidentemente, al combustible de mi vida, ma petite puce, Corale!

[1] Supported by the Programme Alβan, the European Union Programme of High Level Scholarships for Latin America, scholarship No. E07D402742CO.

Contents

CHAPTER 1

Introduction

Contents

"Tout change, quoique pierre"
–Claude Monet.

Résumé étendu (en français). Depuis des centaines d'années l'apparence des objets a été un important sujet d'étude qui a attiré l'attention de plusieurs savants au fil du temps. Dans un premier temps, une intuition de type perceptuel a guidé les chercheurs pour décrire la notion de la "couleur". Par exemple, Aristote (830 av. J.-C.) décrivit le processus de vision comme une altération de l'espace entre l'observateur et sa cible. Cette représentation précoce enferme déjà l'idée qui montre que l'apparence n'est pas une propriété inhérente à l'objet mais une caractéristique qui dépend de l'environnent ou "medium" comme l'a appelé Aristote.

Au cours du XVIII ème siècle le développement d'une rigoureuse discipline scientifique a permit à Newton identifier la lumière comme la source de la perception de la couleur. Dans ces travaux, il explique le phénomène de décomposition de la lumière et il établie les principes qui régissent la réflexion de ses rayons. Ces résultats ont été complémenté par les travaux de Lambert. Une de ces principales contributions est le livre nommé *Photometria* [Lambert 1760], dans lequel l'auteur modélise la relation entre la lumière et le milieu lorsque un rayon traverse des matériaux avec différentes propriétés. Ce rapport est connu sous le nom de la Loi de Bert-Lambert.

Quelques temps après, le célèbre scientifique et écrivain allemand Goethe publia un livre "La Théorie des Couleurs" [von Goethe 1840]. Ce texte n'est pas une théorie dans le sens rigoureuse, mais il décrit graphiquement la nature de la couleur. Son ouvre devient une source d'inspiration pour des artistes et peintres de l'époque. Presque au même temps, en suivant une approche physique au problème, Tomas Young propose la théorie trichromatique. Cette théorie établie que pendant la première étape de la perception de la couleur, dans la vision humaine, la lumière est décomposée par de capteurs primitifs syntonisés a trois différentes longueurs d'onde (les cellules photorécepteurs). Puis, Von Helmholtz et Maxwell ont étendu les idées de Young à partir de nouvelles expériences pour déterminer la correspondance des couleurs.

Du coté artistique, la Renaissance a marqué le début des travaux pour explorer de façon empirique le problème de la perception visuelle. Différents peintres ont étudié le sujet et avec l'arrivé de l'Impressionnisme les artistes ont commencé à analyser rigoureusement le rapport entre la lumière et l'apparence des objets. Un exemple exceptionnel de cette époque est l'étude la lumière fait par Monet dans la cathédral de Rouen. L'artiste a peint une série de tableaux qui correspondent à la façade du bâtiment dans différentes conditions d'illumination. En regardant plusieurs exemples de la série, un observateur peut facilement percevoir toute la richesse d'apparences qui peuvent-être trouvées dans le même objet.

Au milieu du XX$^{\text{ème}}$ siècle le neuroscientifique David Marr a décrit le processus de la vision comme la capacité biologique qui nous permet d'interpréter l'information contenu par un tableau en deux dimensions comme une description en trois dimensions du monde réel. Il propose de modéliser le processus par trois étapes, où la perception de l'apparence fait partie de la seconde étape, un brouillon de la scène (2.5 sketch), qui contient l'information sur les textures et les couleurs. En même temps d'autres chercheurs ont découvert que l'information de la couleur dans les êtres humains est capturé par des photorécepteurs spécialisés appelés cônes localisés à l'intérieur de l'œil. Jusqu'à ce point, il est possible de distinguer les trois principales composantes d'interaction pendant le processus de perception des couleurs : premièrement, les propriétés physiques de l'objet contenu dans la scène. Ensuite, l'illumination et l'environnement qui entoure l'objet et finalement la nature des capteurs utilisés pour attraper la réalité.

Avec l'arrivée de la révolution numérique, l'utilisation d'outils informatiques pour simuler les interactions de ces composants est de plus en plus fréquente. A partir de ces résultats, on a ouvert la porte à des nouveaux domaines comme, par exemple, la réalité augmentée ou la visualisation réaliste. Des systèmes de plus en plus détaillés (et complexes) ont modélisé les comportements des composants du processus de formation de l'image. En plus, la quantité d'information visuelle semble être en état d'augmentation constante. Les appareils photo numériques sont véritablement par tout et les gens peuvent facilement partager ses photos avec quiconque. Ces faits fixent le cadre pour les travaux de recherche présenté dans cette thèse : Quelle sorte d'information peut-être extraite à partir d'une collection d'images ? Pour s'attaquer à cette question, nous proposons différentes approches, tout en soulignant l'importance d'exploiter la richesse d'apparences disponible dans les grandes collections de

photos.

Pour répondre à la question formulé préalablement, nous allons adresser le problème d'analyse de l'information photométrique en utilisant des collections de photos. Pour cela deux différents fils conducteur de la recherche sont proposés. En premier, nous traitons d'une façon empirique le problème de la comparaison l'apparence du ciel, étant donné que cette forme de lumière est la principale source d'éclairage dans les scènes à plein air. Ce sujet est abordé dans le chapitre 3. Le second fil aborde la question en introduisant information géométrique de la scène dans nos calculs. À partir d'une structure 3D obtenue avec des méthodes standard de reconstruction multi-vues et les images appartenant à la collection, nous proposons de modéliser avec un certain dégrée de détail les composants principaux de la formation de l'image : l'illumination pour chaque image, les propriétés de réflectance des objets contenus dans la scène et la réponse radiométrique pour les appareils photo utilisées pendant l'acquisition. Les chapitres 4 et 5 développent ces idées. Il faut ajouter que le chapitre 2 présente un état de l'art pour cadrer nos formulations et que le chapitre 6 résume les contributions résultant de cette recherche et décrite les possibles avenues de recherche pour exploiter les informations photométriques en utilisant des collections d'images.

1.1 Motivation and Context

"Everything changes, even stone"
–Claude Monet.

Analysis of object appearance is a subject that has attracted the attention of many gifted intellectuals over time. Different approaches had been addressed, at the beginning most of them followed some kind of perceptual intuition of the color notion. Aristotle (830 B.C), for example, explained the process of vision as an alteration of the "medium" occupied between the observer and the object. His basic idea of the human vision phenomenon already showed that appearance is not an inherent property of the object, but a characteristic dependent on the environment or "the medium" as Aristotle called it. But it was not until the XVIIIth century that Newton identified the light as the source of the color perception. In his works, Newton explained the light decomposition phenomenon and he established some principles of the light reflection. His results were complemented by the contributions of Lambert. One of his main contributions to this field, a book called *Photometria* [Lambert 1760], establishes the relation between the light and the medium when it travels through materials with different properties. This is knows as the Bert–Lambert Law.

Later on, Goethe, the famous German writer and scientist, published a book named Theory of Colours [von Goethe 1840]. This text is not formally a theory, it tries to describe graphically rather than explain physically the nature of color. However it became the source of inspiration of many painters from that time. Following a physical approach to the problem, Thomas Young proposed the trichromatic theory where he described color vision as the result of the light incoming to three different photoreceptor cells. Von Helmholtz and

Maxwell later expanded Young's ideas using experiments designed to match colors at particular wavelengths.

On the side of the arts, since the Renaissance, different painters have also empirically explored the visual perception process. But it was with the advent of the Impressionism, that the great masters started to analyse rigorously the relation of the illumination and the object appearance. A notable example is the study of light[1] made by Claude Monet on Rouen's Cathedral. In this series, the artist painted more than 30 views of the facade of the building under different illumination conditions. This example illustrates the large variety of appearance that could arise when visualizing the same scene under changing lighting. In recent years (2005) a group of technicians and artists, known under the name of "*Monet aux pixels*", rendered tribute to the work of the painter by projecting images of his ouvrages on the facade of Rouen's cathedral (see figure 1.1). This example presents an extreme case of how the illumination can change the appearance of an object.

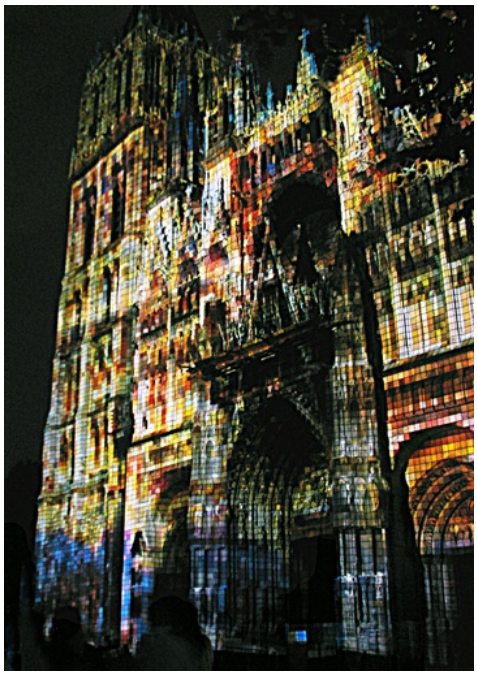


Figure 1.1: Projections of Monet's paintings on the facade of Rouen's Cathedral[2].

During the middle of the XXth century the neuroscientist David Marr formulated the vision process as a series of simple steps in a well defined three-stages pipeline. This biological process allows us to interpret a two dimensional visual array as a three dimensional description of the world. In this pipeline, the color perception could be seen as a part of the second stage, the 2.5D sketch of the scene, where textures and colors are known. At the same time,

[1] web site: `http://www.learn.columbia.edu/monet/swf/`

[2] Credits: Photograph taken by Peter Whelerton (*Darwin70*) via Flickr.

biologists had found that color information in human beings is processed by receptors called cone cells tuned at particular wavelengths. Theses frequencies might vary smoothly from one person to other (which means that color is also a subjective property), but in general a common range of frequencies are widely accepted. This is usually called the tristimulus response of the eye.

At this point, it is possible to distinguish the three main components interacting during the color perception process. For a particular scene we have: Firstly, the properties of the objects contained into scene. Second, the illumination falling into these objects and their environment. And finally, the nature of the sensors used to capture the world.

With the arrival of the digital revolution, the use of computational tools to imitate the interactions of these components has opened the door to new fields as, for example, augmented reality or realistic visualisation. Complex and more detailed models have simulated more and more accurately the behaviors of the actors in the image formation process. Moreover, the amount of information available seems to be increasing continuously. Digital cameras are easily accessible and people want to share their images with family, friends and even strangers. These facts, enclose the main motivation for this work and state a starting point for our research problem: What kind of photometric information can be extracted from a collection of images? To tackle this problem, we explore different approaches. But before getting into details, let us start by describing some of the benefits of working with large image databases and the advantages of having a complete photometric description of the scene.

1.1.1 Large image databases

During the last years, the amount of visual information uploaded to the Internet has increased dramatically. This overgrowth has probably been the consequence of two big milestones that converged at the same time. The first one is associated to the commercial success of digital cameras. Nowadays more than 90% of the mobile phones produced in the world have an embedded camera; also, the market of digital cameras (and their accessories) is more healthy than ever. The second event of this conjunction has been the arrival of a new way of sharing and communicate our experiences using "the friendly face" of the Internet, where any person, even without particular technical skills can express his thoughts. This fact was materialized by the actors of the net by using a new implicit standard called the Web 2.0.

For example, websites like *Flickr*, *Photobucket* or *Getty Images* have specialized their resources in hosting photographies while others have made more emphasis in the social character of the image content (*Facebook* or *TweetPic*). Traditional Internet actors (*Google*, *Bing*) have chosen to follow a path where intelligent image search in large photo collections seems to be the priority. The common point among all these Internet giants is that their servers store billions of images. Among them is common to find large collections of images belonging to a same scene, but taken with different equipments, by different photographers and under diverse illumination conditions. One can find an immensely rich variety of appearances for the same scene. This is the characteristic we want to exploit. Of course, increasing the

Figure 1.2: Websites for upload images and explore photo collections.

number of images brings some disadvantages: higher computing requirements, more noise in the processes and unnormalized ranges of information. These are some of the challenges to overcome in this thesis.

1.1.2 A Photometric Model for the Scene

A scene can be described, in a photometrical sense, if we are able to represent accurately the appearance of the objects immersed in it and their environment from any point of view. Moreover, if the scene is captured by any kind of sensors, a complete photometric description also includes information about the devices used during the acquisition. With this definition in mind, it is possible to set the three main key components for photometrically modeling a scene:

- The illumination filling the environment.
- A description of how the object's materials interact with the light.
- A representation of how the light reflected by the objects is transformed into an image.

But before continuing, let us explain why these components are so important. First of all, knowing the illumination seems to be vital for many detection and classification tasks. Once the illumination is recovered, its effects can be discarded and the processes of recognition, comparison or analysis become more robust and reliable. Secondly, capturing in a simple way an accurate description of how the materials reflects the light is a desirable characteristic for applications that require scene understanding and realistic image synthesis. Usually, a simple way to capture the light reflected by the objects in the scene is by using a photographic

camera as acquisition device. But the way how the final output (an image composed by pixels) is related to the light reflected by the objects can not be deduced in a straightforward form. Lack of knowledge of this relationship is a strong impediment for problems such as color constancy, image–based acquisition or photo–realistic 3D reconstruction.

Joining all the mentioned components in an accurate and efficient representation is the paradigm of appearance modeling in the computer vision and computer graphics worlds. If we have a geometric description of the scene (a 3D model) plus its appearence, one can theoretically generate any virtual image using these components. Applications such as augmented reality, free–view point television or in general realistic modeling are the main targets. In this thesis we study the conduct of these components and their interactions in a particular framework: photo collections containing hundreds of photos of the same scene, captured under unspecified conditions by (not necessarily) amateur photographers.

1.2 Addressed Problems

In this work we address the problem of analysis of the photometric information using photo collections. For this two different lines of research are discussed. In the first one, we treat in an empiric way the problem of comparing the appearance of the sky, given that this is the main source of illumination in outdoor scenes. The second part exploits additional information of the images. Using a 3D model for the objects contained into scene, we propose to recreate the path followed by the light to conceive a digital image. The objects reflect the light coming from an illumination source in the form of photons. After interacting with the object surface material, these particles arrive to the sensor where internal processes relate them to the image intensities. To reach this ambitious goal, this line of research is subdivided in two stages. At the beginning, we emphasize our work on the relation between the image pixels and the amount of energy incoming to the sensor (Radiometric Calibration). Then, we extend our algorithm, by formulating a general model to estimate, in addition to the radiometric calibration, a general illumination model and the reflectance properties of the object.

1.2.1 Image classification from the sky appearance

Understanding natural illumination seems to be the key for the analysis of images taken outdoors. It is easy to prove, when we look at this kind of images that much of the information perceived is due to the illumination coming from the sky. Moreover, when comparing two or more images of the same scene, it becomes easier for the user to deduce the lighting conditions in one or several of them, even if these images do not correspond to the same point of view. In that context, we propose an approach to compare the sky appearance in large image databases. The obtained results give us clues about the nature of the illumination and the way how it is captured in photographs.

1.2.2 Radiometric camera calibration

Radiometric camera calibration consists of decoding the processes implemented in the camera to transform the energy of the photons captured by the sensor into intensities of an image. This process usually involves two stages. First, one must determine the camera response function, *i.e.* the mapping between the incoming light to the sensor and the recorded intensities. Traditional methods involve imaging a calibration target or taking multiple exposure images. The second step consists in recovering the artifacts introduced in the image due to imperfections on the lenses, for example, vignetting or optical fall–off. In this work we will focus on the former aspect, given that the distortion introduced by the imaging system's optics can be assumed to be linear with respect to the incoming light (despite that it may vary spatially over the image).

1.2.3 Joint estimation using a general illumination model

In the framework of a joint optimization, we use 3D models recovered from photo collections with available methods to retrace the complete image formation process. As results we calculate, in addition to the radiometric calibration of the cameras, a model for the global illumination depicted on each image and a basic representation of the surface reflectance properties (albedo). Common methods capture the illumination by inserting a spherical object into the scene, whose surface reflects the environment lighting. Without having access to the scene, this method is not applicable. On the other hand, only one image does not seem to keep enough information to recover an approximation of the illumination. But having a 3D model incorporates new information to the estimation allowing us to approximate the lighting conditions for the environment.

1.3 Structure of the document

This document is written following the chronological order employed for the development of the research. After this short introduction, in chapter 2, we shall explore the process of image formation, making a particular emphasis on the photometric aspects of the pipeline. We shall also describe some issues related with outdoors illumination, and we are going to expose some of the technical tools used in our algorithms.

Chapter 3 presents two approaches to classify images in large databases based on the appearance of the sky. The methods here described gather a series of techniques to segment, model, and compare outdoor images in large photo collections according to the visual aspect of the sky.

Chapters 4 and 5 exploit the geometric information recovered from an unordered photo collection to estimate the parameters that define the photometric representation of the scene: the radiometric calibration, the illumination and the material reflectance properties. Particularly, chapter 4 focuses mainly on the accurate estimation of the radiometric response for

each camera belonging to the photo collection. In this chapter we expose the principles of the method, along with a series of experiments on synthetic and real data. The algorithm described in chapter 5 is based on the same formulation of the method presented in the previous chapter, but this time, the approach uses a more realistic illumination model. We also give details on the robust implementation of the algorithm along with the explanation of some technical issues. Relevant results and conclusions are presented at the end of each chapter.

We conclude this dissertation by summarizing the contributions resulting from this research and by describing what would be the future research avenues to exploit photometric information using image collections.

CHAPTER 2

Background and State of the Art

Contents

Résumé. Dans nos jours, l'accès à des grandes collections de photos est une réalité qui est à portée de main pour n'importe quel utilisateur connectée à la Net. Ce fait est une conséquence directe de la conjonction entre la popularisation des appareils photo numériques et la grande accessibilité à Internet. Les conditionnes particulières ici décrites, ont étés le déclencheur pour capturer l'attention de plusieurs groupes de recherche. En effet, dans les années précédents, scientifiques appartenant au domaine de la vision par ordinateur, ont démarré des importants efforts afin d'exploiter l'information contenue dans ces bases de données. Pour affranchir L'objectif formule dans ce tavail, c'est à dire, l'exploitation de collections d'images pour déduire information de l'apparence de la scène, nous allons consacrer ce chapitre à un état de l'art des modèles de formation d'une image ainsi que des modèles d'apparence du ciel. Nous introduisons quelques outils mathématiques et la notation à utiliser pendant le reste de cette thèse.

The increasing popularization of digital cameras and the broad accessibility to Internet have created the perfect conjunction for the advent of Large Community Photo Collections; thousands of photos are available only one click away. These particular conditions have called the attention of the computer vision community and, in recent years, researchers have started important efforts to exploit the characteristics present in this kind of collections. One

notable example is the seminal work presented by Snavely *et al.* [Snavely 2008] on 3D modeling from internet image collections. This project illustrates perfectly how to exploit the redundant information available in these collections.

Since our objective is to take advantage of the large amount of data available in photo collections to infer appearance information of the scene and the cameras, we present in this chapter a review of the models and techniques that have been used to achieve this goal. For clearness of the document, the mathematical notation used in this work is explained next. We have tried to keep a consistent notation throughout the thesis, however at times the notation has been adapted in the best way to conventions used in the corresponding research literature. Vectors are assumed to be column vectors and are denoted by Roman bold letters. Usually, last letters of the Latin alphabet represent variables (x, y, z) or vectors of variables ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$). A superscript T denotes the transpose of a vector or a matrix. For example, the vector $\mathbf{x}^\mathsf{T}$ is a row vector. Uppercase Sans Serif letters, such as M, correspond to matrices. A row vector with J elements is represented by $[y_1, \cdots, y_J]$ and its correspondent column vector is denoted by $\mathbf{y} = [y_1, \cdots, y_J]^\mathsf{T}$. Greek letters are reserved for variables such as the covariance matrix, angles, and in the context of reflection modeling, the greek letter *rho* (ρ) denotes the albedo.

2.1 The Image Formation Model

The approaches proposed in this thesis are based on the use of images as measurements of physical phenomena. It seems to be a good idea to start by understanding what is an image and how is it formed. Some authors define an image as "...a two dimensional pattern of brightness" [Horn 1986]. This is a valid denotation, but its scope can be extended if we add "A two dimensional pattern of brightness *result of a physical process happening on the 3D space*". In fact, when an image is generated, its intensity values are the consequence of some energy radiated by a physical source. In turn, this physical source emitting an energy is the result of interactions between the illumination source, the reflectance properties and the 3D structure of the object. At the end, the "pattern of brightness" that we call image is the consequence of the interplay of three main instances:

- The illumination.
- The object reflectance properties and its geometry.
- The sensor properties.

In the next subsections we will detail each of theses components and we will describe some of the mathematical models frequently used to simulate their physical behaviour.

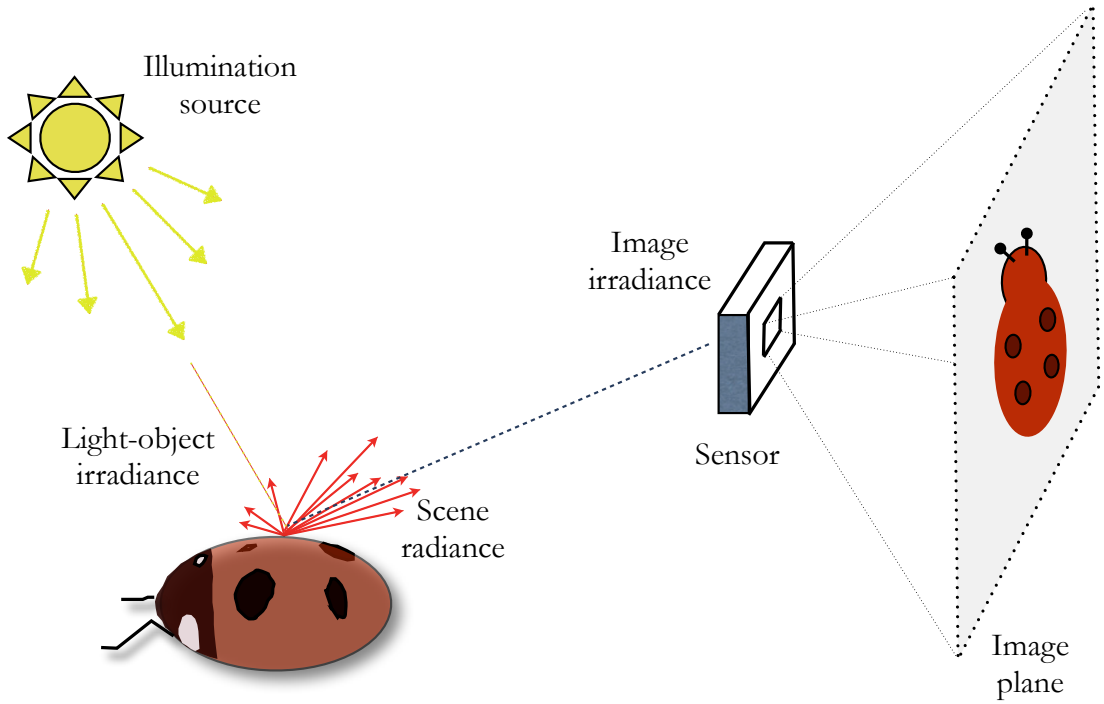


Figure 2.1: Example of the image formation process

2.1.1 The Image: a 2D Projection of a 3D World

To uncover this complex process, there are two key aspects that should be studied: The first one is related to the position of the points in the 3D space and their corresponding positions on the 2D image plane. We call this the **image geometric projection**. The second aspect determines how bright the image of some point on the surface will be. In this part we explore how the object's appearance is associated to the pixel value represented on the image. The latter process is called **image color formation**.

Image Geometric Projection. The Computer Vision and Photogrammetry communities have explored deeply the process of image formation based on the use of projective geometry. Important works have compiled the main achievements reached in this subject and they are excellent sources of reference [Faugeras 1993, Hartley 2003, Szeliski 2009]. The idea is to find a mapping of the 3D point position into the 2D grid describing the image. Frequently, the used model makes a good balance between the complexity of the equations and the approximation of the physical phenomena. This mapping from $\Re^3$ to $\Re^2$ is dependent on two sets of parameters that describe the geometry of the camera: the **extrinsic parameters** representing the pose (rotation and position) of the camera relative to the real world and the **intrinsic parameters** expressing the internal properties of the camera, usually denoted by the focal distance (f), the pixel density on the sensor (k) and the size of the sensor.

The estimation of the intrinsic parameters is known as geometric calibration of the camera and can be done pointing the device to a specific target [Bouguet 1998]. Nevertheless, the computation of the intrinsic parameters is associated to the extrinsic parameters and, as a consequence, the standard estimation process usually finds both sets of values. The process of calibration consists, roughly speaking, in solving a system of equations formed by the projection of some known points visible on the target to the corresponding points reproduced in the pixels of the image. Semiautomatic and automatic generalizations of this idea had been developed using know structures on the scene [Sturm 1999, Criminisi 2000, Liebowitz 1998].

Most of the mentioned processes require an automatic correspondence of the visible regions through multiple views. This goal can be reached by extracting and matching a group of particular interest points in the concerned images. Approaches for discover interest points are abundant, *e.g.* [Moravec 1980, Harris 1988, Shi 1994], and improvements on the robustness of algorithms for matching this kind of features are constantly published [Fischler 1981, Nister 2006, Weiss 2009, Muja 2009]. Moreover, with the arrival of robust techniques —invariant to scale and illumination— the process of feature matching between two or more images could be almost completely automatized [Lowe 2004, Bay 2008]. This fact and the inclusion of the epipolar geometry, open the door to new techniques of autocalibration using multiple images.

Working on the same direction, researchers developed other important applications. For example, some algorithms have focused their efforts on the joint estimation of the camera parameters and a sparse set of points describing the 3D structure of the scene. Initially called **Structure from Motion** (SfM) algorithms (because of the use of one single moving camera) they try to recover the 3D position of a spare set of points on the scene and also the trajectory of the dynamic camera. In our days, the expression SfM has been generalized not only to images acquired with a single camera, but also to the scenes acquired with different sensors from different positions. In the seminal work of Snavely *et al.* [Snavely 2008], authors used SfM algorithms to estimate the camera's pose and to render thousands of images from an photo collection in an intuitive and pleasant way. This process is the backbone of the well known web application *PhotoSynth*[1]. The presence of outliers is common during this process, mainly when the control over the scene and acquisition is minimum. Robust techniques are usually implemented.

Usually, in order to refine the results, a final optimization stage is added to the process. This last part of the pipeline is known in literature under the name of (sparse) **Bundle Adjustment** [Triggs 2000, Hartley 2003]. It consists of an optimization step, frequently solved using the Levenberg-Marquardt algorithm [Marquardt 1963]. When the number of input images is large, it requires often a strong processing power and its implementation must benefit of the sparse nature of the problem. A very efficient implementation of the algorithm profiting of this characteristic can be found in [Lourakis 2009]. Figure 2.2 shows the results of a sparse reconstruction using images downloaded from the internet and the tools described in

[1] http://photosynth.net/

Snavely's work [Snavely 2008].

Another of the notable applications issue from the rigorous study of the camera-scene geometry is called **multiple view stereo reconstruction** (for a survey please refer to [Seitz 2006]). The goal of these algorithms is to reconstruct a complete 3D model (geometry and appearance) from a set of images taken from known camera viewpoints. In this case photometric information plays an important role. Most of the works rely their success on the set up of controlled conditions and a careful acquisition. Given that the camera parameters are known, the optimal solution to this problem is usually linked to a good photometric measure and the way how the algorithm handles ambiguities and occlusions. Successful approaches can be found, for example, in [Gargallo 2008, Delaunoy 2008, Esteban 2008].

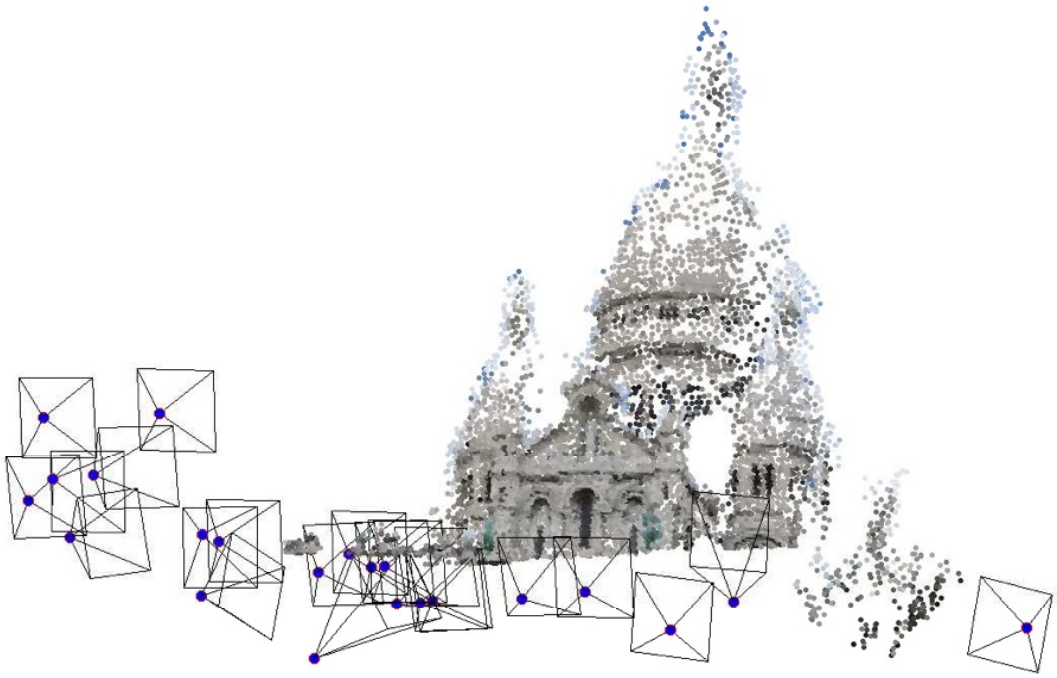


Figure 2.2: Example of a 3D reconstruction using [Snavely 2008].

Recently multi–view stereo reconstruction has been successfully tested using unordered and unstructured collections of photos. Furukawa *et al.* [Furukawa 2010, Furukawa 2009] implement a technique for clustering images taken from close camera positions and pointing to similar parts of the scene. Authors produce a dense set of points representing the 3D structure of the scene. The common point of all the mentioned approaches is the use of a robust measure of the coherence between the projection of a 3D point and the value given by the image. This function is called the **photometric consistency measure**.

Estimation of perfect models for the appearance and, at the same time, the geometric structure is the paradigm of multi-view stereo reconstruction algorithms. However, this

"utopia" is impossible to accomplish without additional hypothesis. As we shall illustrate later, the image color formation is a process also dependent on the object geometry and, as consequence, its appearance is linked to its morphology. This 2-way relation leads us to the classical chicken-egg problem: To estimate the 3D structure it seems essential to calculate the reflectance properties of the surface and, in turn, to estimate object's appearance we need to know its 3D structure! Luckily, as shown in the previously mentioned works, particular hypothesis on the appearance of the object material have been good enough to produce accurate estimations of the 3D geometry of the scene.

Image Color Formation. The second aspect to be studied during the generation of the image is the process that determines how bright a particular pixel is. Traditionally, this subject has been explored by the Image Processing community [González 2008, Pratt 2001], but different knowledge areas have also made important contributions. For example, from the physical point of view, the light is seen as an electromagnetic radiation flowing in the space. Radiometry and Photometry try to characterize the "amount" of this flow [Ishimaru 1991]. Computer Graphics has also participated on the decoding of the image formation process, by proposing novel ways to estimate object appearance in real conditions [Sillion 1994, Debevec 1996, Dorsey 2007]. As already mentioned, the pixel intensities stored in a digital image are the result of the interaction of light, object's photometric and geometric properties and sensor characteristics.

Before going further we want to define some terms commonly used in Photometry to describe the image formation process. First, we define the amount of light falling onto a defined surface. This measure is called **irradiance** and it is expressed in $[W/m^2]$. In the scheme shown in figure 2.1 there are two types of irradiance: the light–object irradiance, *i.e.* the amount of light falling on the object, and the **image irradiance** that is the amount of light reaching the area limited by the sensor. For our analysis, we shall make emphasis of the former. Additionally, we introduce the term **radiance**. It is defined as the light emitted from a surface per unit of area in a specific direction [Zickler 2009]. This direction is given in terms of a solid angle. It is measured as the power emitted per unit of projected area per unit of solid angle $[W/(m^2 \cdot \mathrm{sr}^{-1})]$. The **scene radiance** is completely characterized by the reflectance model of the material composing the surface. There exist different reflectance models. High-detailed representations can simulate most of the effects happening in real world scenes (scattering, wavelength dependence, interreflections, etc.), but the price to pay, in terms of the computational cost and the complexity of the calculations, is usually very high. Frequently, the reflectance model is well described by the Bidirectional Reflectance–Distribution function (BRDF), a model that relates the incoming irradiance given an illumination source to the outgoing radiance of the scene. The function that represents the scene radiance for every direction in every point (and for every wavelength) is known as the **Plenoptic function** [Adelson 1991]. Images can be seen as samples of the Plenoptic function. This function represents completely the photometric properties of the scene and it is frequently used in augmented reality applications to create realistic environments.

Image irradiance, scene radiance and all the other definitions made for the moment are variables depending on the wavelength of the light. Sensors of modern cameras have filters to capture the spectral distribution of the image irradiance. The trichromatic components output from the camera sensor are actually a measure of the amount of light corresponding to a wavelength passing through the filters. Assuming that camera sensitivities can be modeled by a delta Dirac function [Finlayson 2006], $Q_k(\lambda) = q_k\delta(\lambda - \lambda_k)$, where $k = \{R, G, B\}$ depends on the filter used, then the image irradiance E_k for an element at the surface position $\mathbf{x} = [x_1, x_2, x_3]^\mathsf{T}$ is expressed as a function of the light wavelength:

$$E_k(\lambda, \mathbf{x}) = \int L(\lambda, \mathbf{x})S(\lambda, \mathbf{x})Q_k(\lambda)d\lambda \tag{2.1}$$

$$E_k(\mathbf{x}) = L(\lambda_k, \mathbf{x})S(\lambda_k, \mathbf{x})q_k \ , \tag{2.2}$$

where $L(\lambda, \mathbf{x})$ models the spectral power distribution (SPD) of the illumination falling on the surface position $\mathbf{x}$ and $S(\lambda, \mathbf{x})$ the bidirectional reflectance–distribution function (BRDF) of the object in those coordinates. According to the equation (2.2), both functions ($L(\lambda_k, \mathbf{x})$, $S(\lambda_k, \mathbf{x})$) are dependent on a fixed light wavelength λ_k. Henceforward, we shall refer to these functions $L_k(\mathbf{x})$ and $S_k(\mathbf{x})$ by knowing implicitly their dependence to a particular light wavelength.

Finally to obtain the image brightness, our image formation model establishes a non–linear function for the image irradiance $E_k(\mathbf{x})$. This function is called the camera response function (CRF). In the next subsections we will explore in some detail each of the components of image formation and their relation with the environment and the object appearance.

2.1.2 Surface Reflectance Models

In equation (2.1) we describe the image irradiance as a function dependent on the light wavelength and the position. Now we consider the process of formation of the scene irradiance, *i.e.* the light reflected by the object in all directions. Let us ignore for the moment the influence of the light wavelength λ and consider the process as "independent" of this parameter. The scene radiance is the result of the incoming light and the object reflectance properties. These properties are characterized by the appearance model of the surface, also know as the bidirectional reflectance–distribution function (BRDF) [Nicodemus 1992]. Formally, the BRDF is the ratio between the scene radiance (L_o) and the incoming irradiance I_i. It is usually represented as a 4D–function written in terms of the incident light angles $(\theta_i, \phi_i,)$ and outgoing radiance angles (θ_o, ϕ_o) with respect to the normal of the surface in the position $\mathbf{x}$. Then,

$$S(\mathbf{x}) = S(\theta_i, \phi_i, \theta_o, \phi_o) = \frac{dL_o(\theta_o, \phi_o)}{dI_i(\theta_i, \phi_i)} \ . \tag{2.3}$$

The BRDF has two important properties:

- The first one is the **energy conservation**. All incident light must be reflected or absorbed by the surface material and no "extra" light is created during the process. This

is expressed in mathematical notation by integrating the scene radiance over all the outgoing directions (Ω), and scaling the results by the cosine term results of the foreshortening: $\int_\Omega S(\mathbf{x}) \cos\theta_o d\theta_i d\phi_o \leq 1$.

- The second property, also known as the **Helmholtz reciprocity**, states that the BRDF must be unchanged when the incidence and outgoing angles are swapped.

Approximate–analytical models for the BRDF are usually described using two components: one for the diffuse reflection and a second one for the specular contributions. Among the proposed models one can find the Phong model designed to imitate the appearance of glossy surfaces [Phong 1975], the Lafortune model used to fit measured BRDF data to analytic functions [Lafortune 1997], the Ward model which includes a specular term shaped by a Gaussian function [Ward 1992] and the Cook-Torrance BRDF model that mimics the specular reflections by adding "microfacets" —*i.e.* numerous tiny mirror-reflective components— to the diffuse term of the surface [Cook 1981]. Although all these models try to simulate real surfaces with more or less details not one can be suitable for all kinds of scenes. Also, the number of parameters when dealing with specular reflections increases dramatically because the scene radiance is dependent also on the incoming light angles.

On the other hand, the Lambertian BRDF is characterized by the absence of specular regions on the modeled surface. It is the simplest way to represent surface reflectance, but it captures the essence of many real world materials. Moreover, the specular particularities present in some surfaces are most of the time reduced to particular small regions and they can be treated as outliers of the Lambertian BRDF using robust algorithms. The model is just a constant ρ called the **albedo**, divided by π to guarantee the "energy conservative" property when integrating over a semi sphere:

$$S = \frac{\rho}{\pi} \ . \tag{2.4}$$

The lambertian BRDF is included in the image irradiance equation (2.1) (which has the "incident cosine law" implicitly included). The resulting equation for the irradiance E of the j^{th} point localized in coordinates $\mathbf{x}$ is the following:

$$E_{jk} = \frac{\rho_j}{\pi} \cdot q_k \int_\Omega L_k(\theta_i, \phi_i) \cos\theta_i d\Omega \ , \tag{2.5}$$

where Ω denotes the hemisphere of all possible incoming light directions. We observe also, that in the case of Lambertian BRDF, the irradiance E_{jk} is only dependent on the incoming angles (θ_i, ϕ_i). That means that the position of the observer (or camera) does not influence the magnitud of the image irradiance.

2.1.3 Illumination Models

Generating new views of a 3D model with Computer Graphics tools usually requires knowledge of the nature of the illumination present on the scene. Inside this framework, researchers

have proposed models to simulate light behavior in different ways, *e.g.* [Schlick 1994, Heckbert 1992]. There exists a wide range of models going from simple and computationally inexpensive representations to high-detailed models that imitate almost perfectly the properties of the light in real world scenarios. When an illumination model must be chosen for a specific application, the compromises between accuracy and complexity should be carefully taken into account. We describe in this section the two illumination models used in chapters 4 and 5.

Directional light source. We will start with a simple model. Assuming a scene where direct illumination is dominant, *i.e.* the light arrives to the object straight from the light source. For the case of a Lambertian surface, the scene radiance is the same in all the directions and image irradiance does not depend on the position of the viewer. Then, equation (2.5) is expressed as the dot product of the directional light source and the normal describing the surface at point j. The result of this operation is called the *diffuse* term of the rendering equation. Additionally, we include a constant term to model the contribution of the ambient illumination to the image irradiance. Thus, the image irradiance for a surface point j with coordinates $\mathbf{x}_j = [x_{1j}, x_{2j}, x_{3j}]^\mathsf{T}$ illuminated by the i^{th} light source $L_{ki}(\theta_i, \phi_i, \beta_{ki}, \mu_{ki})$ is computed by:

$$E_{kji} = \frac{\rho_j}{\pi} \cdot q_k(\mathbf{n}_j{}^\mathsf{T}\mathbf{l}_{ki} + \mu_{ki}) \quad , \tag{2.6}$$

where,

$$\mathbf{l}_{ki} = \begin{bmatrix}\beta_{ki}\cos\theta_i\cos\phi_i & \beta_{ki}\sin\theta_i\cos\phi_i & \beta_{ki}\sin\phi_i\end{bmatrix}^\mathsf{T} \quad .$$

Variables (θ_i,ϕ_i) describe the direction and (β_{ki}) the "strength" of the directional light source while μ_i corresponds to the ambient term. ρ_j is the albedo and $\mathbf{n}_j$ is the normal of the surface point $\mathbf{x}_j$. Following the notations introduced by Luong *et al.* [Luong 2002], we can express equation (2.6) using a dot product of the vectors $\mathbf{L}_{ki} = \begin{bmatrix}\mathbf{l_{ki}} & \mu_i\end{bmatrix}^\mathsf{T}$ and $\mathbf{N}_j = \begin{bmatrix}n_{j1} & n_{j2} & n_{j3} & 1\end{bmatrix}^\mathsf{T}$. Then, the image irradiance of a surface element j, observed under light i is given by:

$$E_{kji} = \frac{\rho_j}{\pi} \cdot q_k\left(\mathbf{N}_j^\mathsf{T}\mathbf{L}_{ki}\right) \quad , \tag{2.7}$$

where q_k and k are defined per channel.

Environment Maps. Another common representation for the illumination, usually employed in Computer Graphics applications is known in the literature as reflection mapping or environment maps. Assuming a distant illumination, the lighting is represented as a function of direction related to the scene. In practice, environment maps can be acquired by placing a mirror-reflectance sphere on the scene when images are taken. The sphere simply reflects the incident lighting. This approach is implemented by Debevec *et al.* [Debevec 2004, Debevec 1996] to successfully render models with realistic appearance. Actually Blinn and Newell [Blinn 1976] were the first who used this method to efficiently find reflections of distant objects. Then, the technique was generalized by Miller and Hoffman [Miller 1984]

who proposed to photograph a mirror sphere to acquire a real environment illumination. In [Heidrich 1998], Heidrich and Seidel introduced the use of environment maps with general BRDFs, however the computational cost of the rendering process was expensive. The almost simultaneous publication of the works developed by Basri and Jacobs [Basri 2003] and by Ramamoorthi and Hanrahan [Ramamoorthi 2001b] marked a milestone on the rendering of realistic computer–generated models. In their works, they proposed first to use spherical harmonics [Stein 1971] to represent the illumination of the scene as a "lighting function". Then, using a frequency framework, the image irradiance is computed from the convolution of this "lighting function" and the reflectance function describing the surface. Using these results and under particular circumstances (a Lambertian BRDF and a convex surface) most of the energy generated by the image irradiance can be recovered using only some coefficients of the spherical harmonics describing the lighting function. Later, in section 5.1 we will detail how the global illumination and a Lambertian BRDF are computed using the spherical harmonics.

2.1.4 The Camera Response Function

In equation (2.1) we propose a formula to represent image irradiance incoming to a particular color filter on the camera sensor. However, these variables do not correspond exactly to the values stored on the RGB digital image. Discounting alterations due to the lenses aberrations (vignetting, depth of field, etc.) or other artifacts as, dead pixels, actual camera manufacturers try to design specific responses to different input signals in order to match the way how appearance is perceived by the human vision system. The way how these processes alter the image signal is kept as trade secret by camera makers and little information can be found in this regard.

In the present work we define the Camera Response Function (CRF) as the relation between the image irradiance($E_k(\mathbf{x})$) and the image brightness ($B_k(\mathbf{x})$). This relation is given for the color channel k. Assuming that all the pixels belonging to an image have the same CRF and omitting the subscript denoting the light wavelength, the CRF is expressed as a non–linear function f:

$$B = f(E) \ . \tag{2.8}$$

A classical example of the CRF is the *gamma correction curve*, a nonlinear operation used in the video industry to code and decode the luminance represented by imaging systems. The figure 2.3 depicts the gamma correction curve reprsented as the relation between the irradiance and the intensity on an image: $B = f(E) = E^{\gamma}$, for commonly used values of γ.

Estimating the function f is commonly referred in literature as **Radiometric Calibration**. There exist different approaches to accomplish this goal. Two categories can be distinguished: *active* and *passive* methods. The active recovery of CRF's gathers all the methods that require physical presence during the acquisition process. Common approaches include using either specific calibration objects (color charts such as the popular Gretag-

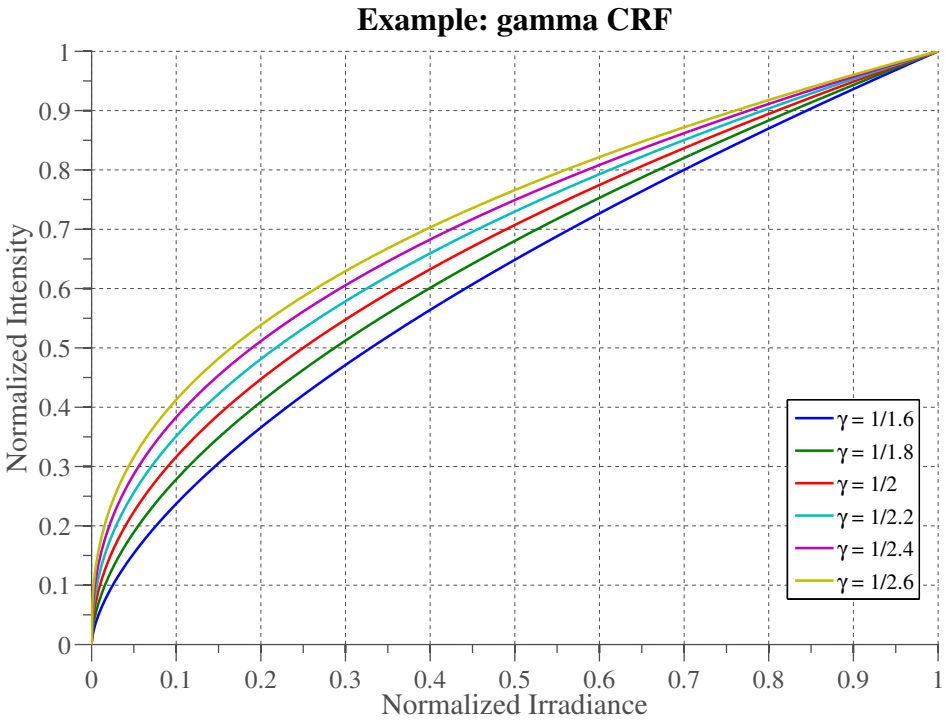


Figure 2.3: CRF example: The gamma correction function. We show the plots for different vauels of γ parameter

Macbeth ColorChecker®[2] [Chang 1996, Ilie 2005] or images acquired by a static camera but with varying exposure times [Debevec 1997, Mitsunaga 1999]. A second category includes approaches where the physical presence on the scene is not necessary. This kind of algorithm seeks image characteristics that reflect the non-linearity produced by the camera response function. Low level representations as edges in regions where color changes drastically [Lin 2004], [Lin 2005] or geometric invariants [Ng 2007] allow to extrapolate aproximations of the CRF. Other artifacts common in digital images such as noise can also be used to estimate the radiometric response of a camera [Takamatsu 2008]. In [Kuthirummal 2008], Kuthirummal *et al.* found priors for statistics on large photo collections. These priors allow them to calculate the response functions of generic camera models, assuming that all instances of one camera model have the same properties and that many images taken by the camera to be calibrated, are available. Personalized devices, *e.g.* , cameras with interchangeable lenses, can not be modeled using this method.

It is clear that having some samples of B and E is not enough for estimating the function f. Having some a priori information about the possible CRFs is necessary. Some parametric models have been proposed. Following the behaviour of the intensity signals on cathodic ray displays, one of the first attempts of modeling the CRF consisted of using a *gamma*

[2] http://www.ae5d.com/macbeth.html, http://www.xrite.com

function [Mann 1995]. Debevec and Malik proposed an approach using multiple images of a static scene taken with different exposures. To recover the CRF they found the best–fitting function in a least–squares sense using a logarithmic space. Mitsunga and Nayar's work modeled the inverse CRF as polynomials of a particular order [Mitsunaga 1999], while Chakrabarti *et al.* propose to come back to the polynomial model adding a matrix for linking the information available on the 3 color channels [Chakrabarti 2009].

Among the works published for modeling the CRF, one that has gained important notoriety is the analysis presented by Grossberg and Nayar [Grossberg 2003, Grossberg 2004]. Authors propose to model the space of CRFs using a database of real world functions. They limit the space of possible CRFs using the following simple assumptions:

- The response function f is the same at each pixel. This assumption is reasonable given that the changes in the behaviour of the CCD from one region to another are minimal, however it might be problematic when handling potential residual vignetting effects in images.
- The range of output intensities (B) generated by the camera is bounded by a minimum and a maximum intensity.
- The CRFs are monotonic functions. In practice this assumption is plausible and coincides with the properties shown in real world CRFs.

The main contribution of Grossberg and Nayar's work is an empirical model of the CRF. This model can be expressed as a linear combination of a basis and some coefficients. The mentioned basis is the result of computing a principal component analysis (PCA) over a set of 201 real–world response functions. Then, a particular CRF can be represented by the coefficients $\mathbf{w} = [w_1, w_2, \cdots, w_n]^\mathsf{T}$ and the basis $H(E) = \{h_0(E), h_1(E), \cdots, h_n(E)\}$:

$$f(E) = h_0(E) + \sum_{n=1}^{N} w_n h_n(E) \quad . \tag{2.9}$$

In practice, the average CRF ($h_0(E)$) and the principal components can, for example, be represented as lookup tables or polynomials fitted to these. Without loss of generality, we use the latter in our work. The main difference between the two representations is that the polynomial representation requires less parameters, while the lookup tables needs to allocate a large amount of memory (depending on the precision demanded by a specific application). Besides this issue, the formulas underlying the methods proposed in this thesis would be strictly analogous in the case of lookup tables. The basis CRF's are thus represented as polynomials of degree D:

$$h_n(E) = \sum_{d=0}^{D} c_{nd} E^d \quad . \tag{2.10}$$

Note that according to [Grossberg 2004], the values h_n are expressed relative to normalized brightness and irradiance values, such that $c_{n0} = 0$ and $\sum c_{nd} = 1$ for all $n = 1 \cdots N$. The

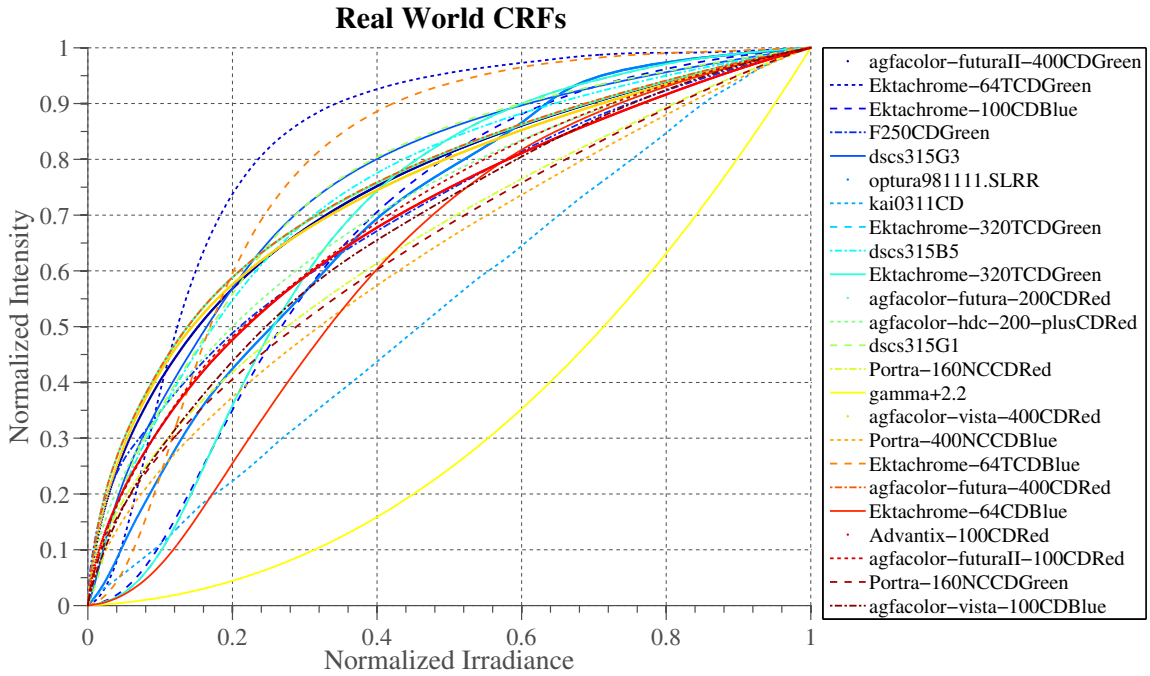


Figure 2.4: Real world CRFs

degree of the polynomials was chosen for an adequate representation of the curves that form the basis, which was obtained with $D = 9$.

Figure 2.4 shows some of the real world CRF's used by Grossberg and Nayar [Grossberg 2004] to create the basis of the empirical model. Figure 2.5 presents the first three functions of the basis h_0, h_1, h_2 and their approximations using polynomials of degree 9. The curves described by polynomials are close together to the original curves and their difference is almost imperceptible.

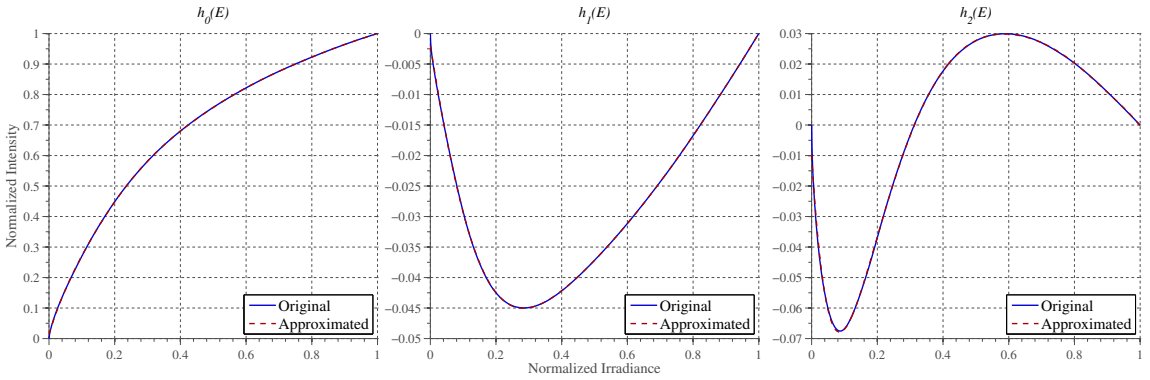


Figure 2.5: First the component of the basis to model any CRF and their approximation using 9-degree polynomials.

The most standard way to capture the CRF is to use a color chart. This instrument gives us known values of the irradiance for a particular illumination and surface. However the use of this element has raised some controversy. In section 4.3 we will describe the protocol used to accomplish this task using images from our own database. Also we will also detail our own implementation of the algorithms proposed by Lin *et al.* [Lin 2004] and Grossberg and Nayar [Grossberg 2002].

2.2 Models for sky appearance

In the first section of this chapter (2.1) we presented a model for the image formation using two alternatives for the illumination representation. Because of its primordial role in outdoors scenes, here we shall describe some popular models to represent the illumination coming from the sky. Following the taxonomy introduced by Lalonde [Lalonde 2011], outdoor illumination is typically modeled by representations that can fall in one of these three categories:

- Physically-based models. This kind of representation consists of parametric models where the parameters have a direct relation with physical characteristics of the environment. For example, in [Love 1997] the author presents a detailed study of the sky light illumination, decomposing the total contribution in two components: direct irradiance from the sun and irradiance from the sky. Each of these sources of illumination are treated separately and some models with parameters like sun position, atmospheric scattering terms, etc. are presented. Another important contribution is the work of Slater and Healey [Slater 1998] where the authors express the spectral irradiance falling into the ground as a function of the sun and sky radiances. They associate the spectral properties of outdoor illumination sources to atmospheric conditions and scene geometry. Their model also includes parameters for modeling the solar and scattered radiations and the effects of atmospheric gases and aerosols. In [Igawa 2004], authors propose a complete physical model for sky radiance and luminance distribution. Their contribution includes representations for clear skies and overcast skies, as well. The method introduced by Ward [Ward 1994] is very appreciated for rendering architectural scenes in the Computer Graphics community. In his approach, the position of the sun is an important factor to calculate the irradiance of the scene. In [Lalonde 2008] authors introduce a sky model as a function of the camera pose and intrinsic parameters. Their work is based on the model proposed by Perez *et al.* [Perez 1993] where the luminance of a point in the sky is a function of its elevation and relative orientation with respect to the sun. The representation proposed in [Lalonde 2011] proposes a mixed model where a physical approach for the sky is combined with some clues extracted from the image, using statistical information inferred from image databases.

- Reflection mapping (*cf.* 2.1.3). Environment map representations were already intro-

duced in section 2.1.3. Unlike physical–based models, this representation does not require the estimation of some parameters because it stores directly the measured incoming light. The main disadvantage, besides the large memory requirements, is in principle, the necessity of physical access to the scene.

- Statistical models. This category, also known as "data-driven" models, uses a priori information extracted from samples or large databases to predict and interpret the appearance of the sky. In [Judd 1964] —considered as a classical work— authors develop a series of experiments using 622 samples of daylight radiance. As a result, they propose to approximate the observation set by using a linear combination of three fixed functions. The use of image sequences of static cameras has been also proposed for the estimation of the components forming the illumination radiated by the sky [Sunkavalli 2008]. Recently, Romeiro *et al.* [Romeiro 2010] addressed the problem of surface reflectance estimation under unknown illumination. In their work, they assign a Bayesian probability to the illumination. This Bayesian framework is calculated a priori using a dataset of possible lightings represented as environment maps.

2.3 Estimation Methods

We briefly describe some of the mathematical tools used in chapters 4 and 5 to estimate the model parameters. The estimation problem is formulated in a least–squares optimization framework. In those cases the goal is to minimize the square of the difference between a vector of samples ($\mathbf{x} = [x_1, x_2, \cdots, x_M]^\mathsf{T}$) and the values predicted by a non–linear model f, where $f : \Re^N \rightarrow \Re^M$ parameterized by the unknown values $\Omega = [\omega_1, \omega_2, \cdots, \omega_N]^\mathsf{T}$.

$$\min_{\Omega} \sum_{i=1}^{M} \left(\mathbf{x}_i - f_i(\Omega)\right)^2 \quad . \tag{2.11}$$

There exist different approaches to compute the values of Ω. They consist of iterative processes exploiting information given by the Jacobian of the function. This information is used to choose the direction of change for the parameters in the next iteration. Among the algorithms available in literature, we can find the **steepest descent** algorithm, the **Gauss–Newton** method or the **Levenberg–Marquardt algorithm**. The last of the three–mentioned method has become extremely popular in computer vision applications. Its success can be explained because it frequently finds a good trade–off between accuracy and speed. In fact, the LM algorithm can be thought of as a combination of the two other mentioned methods: when the current solution is far from the correct one, the algorithm acts like the steepest descent method. Once the estimated parameters are getting closer to the solution, the algorithm becomes a Gauss-Newton method and one can trust to the change direction estimated by the method. The LM algorithm was introduced by Levenberg in [Levenberg 1944] and

then rediscovered by Marquardt in [Marquardt 1963]. For more comprehensive treatments, a full–detailed description of the method can be found in Moré's work [Moré 1978].

We seek to minimize the squared distance $\varepsilon^{\mathsf{T}}\varepsilon$ where $\varepsilon = \mathbf{x} - f(\Omega)$. The strategy to find the values of Ω, also called the *parameter vector*, is based on the linear approximation of f given by the Taylor series:

$$f(\Omega + \delta_\Omega) \approx f(\Omega) + \mathsf{J}\delta_\Omega \ . \tag{2.12}$$

where J is the Jacobian matrix of the function $f(\Omega)$. Then, the iterative process consists in finding a series of vectors $(\Omega_0, \Omega_1, \cdots, \Omega_{\text{opt}})$, starting by an initial guess (Ω_0) and ending when $f(\Omega_{\text{opt}})$ is a local minimum. At each step, a direction (and a magnitude) of change is represented by δ_Ω. Thus, the direction of change is found by formulating the normal equations derived from the following linear least–squares problem:

$$||\mathbf{x} - f(\Omega + \delta_\Omega)|| \approx ||\mathbf{x} - f(\Omega) + \mathsf{J}\delta_\Omega|| = ||\varepsilon - \mathsf{J}\delta_\Omega|| \ ,$$

where, δ_Ω is the solution of $\mathsf{J}^{\mathsf{T}}(\mathsf{J}\delta_\Omega - \varepsilon) = 0$. Rearranging this result we arrive at the normal equations of the problem:

$$\mathsf{J}^{\mathsf{T}}\mathsf{J}\delta_\Omega = \mathsf{J}^{\mathsf{T}}\varepsilon \tag{2.13}$$

For the case of the Levenberg-Marquardt algorithm, the system to be solved is actually a slight variation of the system presented in equation (2.13):

$$\left(\mathsf{J}^{\mathsf{T}}\mathsf{J} + \lambda\mathbf{I}\right)\delta_\Omega = \mathsf{J}^{\mathsf{T}}\varepsilon \ , \tag{2.14}$$

where the identity matrix is denoted by $\mathbf{I}$ and λ is a non–negative scalar for each iteration referred to as the *damping term*. For each iteration the parameter vector is updated by $\Omega_{n+1} = \Omega_n + \delta_{\Omega_n}$. If the new Ω_{n+1} leads to a reduction of the error ε the process passes to the next iteration with a decreased damping term. Otherwise, the systems stays at the same iteration and recomputes the normal equations with an increasing value of the damping term. The algorithm changes to the next iteration when a value of δ_Ω that decreases ε is found.

To judge the convergence of this process, one uses typically one of the following criteria, or a combination there of:

- The value $||\varepsilon^{\mathsf{T}}\varepsilon||$ is less than a fixed threshold.
- The change in the search direction δ_Ω is less than a predetermined value.
- The maximum number of iterations is reached.

Other variations of the LM algorithm take into account the distribution of the sample vector $\mathbf{x}$. These methods include the covariance matrix $\Sigma_{\mathbf{x}}$ into the normal equations. Some algorithms derive directly from the LM method. They try to generalize these results by adapting the search region to the analytic form described by $\varepsilon^{\mathsf{T}}\varepsilon$. These methods are found in literature under the name of **trust–region** algorithms [Yuan 2000]. A notable implementation of the LM algorithm on the framework of Sparse Bundle Adjustment (SBA) is presented by

Lourakis and Argyros [Lourakis 2009]. Later, in section 5.2.2 we will compare the SBA implementation proposed by the named researchers and the approach introduced in this thesis aiming at the joint estimation of photometric properties and the camera response functions.

2.4 Classification and Comparing Metrics

The chapter 3 of this thesis develops a method to select similar images based on their sky's appearance. In the present section we describe and formalize some of the tools used to achieve the mentioned goal.

2.4.1 Mixture of Gaussian Distributions and Expectation–Maximization (EM)

The mixture of Gaussian distribution consists of a probabilistic model used to represent subpopulations that are normally distributed within an overall population. When modeling a dataset, this statistical representation is also known as Gaussian mixture model (GMM). In simple terms, the GMM can be seen as the superposition of single distributions, in this case Gaussians. Using a sufficient number of Gaussian distributions, samples of a particular dataset can be modeled with arbitrary accuracy, even if they represent extremely complex distributions. Aside from the number of distributions composing the mixture, the parameters that govern a GMM are the means and the covariances as well as the mixture coefficients (*i.e.* the contribution of each of the Gaussians). The probabilistic distribution of the data $\mathbf{x}$ is formed by a linear combination of K Gaussians according to the following equation:

$$M(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x} \mid \mu_k, \Sigma_k\right) \quad , \tag{2.15}$$

where $\mathcal{N}\left(\mathbf{x} \mid \mu_k, \Sigma_k\right)$ denotes a component of the mixture —a single Gaussian distribution— with mean μ_k and covariance Σ_k.

Let us consider a set of samples as an incomplete dataset of the phenomenon we want to model. We can use the EM algorithm to find out the parameters that describe the model. Initially proposed by Dempster *et al.* [Dempster 1977] this algorithm has become powerful statistical tool for modeling and clustering datasets. It has been used also to fit the data into a mixture of Gaussian distributions. The method is extensively studied in [Bilmes 1998, Bishop 2006, McLachlan 1997], then here we shall present only a brief description of the technique.

The main idea is to find the maximum likelihood solution for the parameters μ_k, Σ_k given N samples of the distribution. These samples are stored in row order in the matrix X. Suppose we call Ω all the parameters $\{\mu_k, \Sigma_k\}$. Thus, we search the estimation of $\hat{\Omega}$ that

maximises the logarithm of the likelihood function ($\mathcal{L}$):

$$p(\mathsf{X} \mid \Omega) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \Omega) = \mathcal{L}(\Omega \mid \mathsf{X}) \tag{2.16}$$

$$\hat{\Omega} = \max_{\Omega} \log \mathcal{L}(\Omega \mid \mathsf{X}) \tag{2.17}$$

When the parameters Ω describe a single Gaussian distribution the solution that maximises the likelihood can be easily found, by setting the derivative of equation (2.17) to zero. In the case of GMM the solution for $\hat{\Omega}$ is not evident, at least in terms of analytical equations. The Expectation–Maximisation (EM) algorithm proposes optimal–local solutions based on the assumption that the samples are incomplete and that there exist some hidden (or latent) variables. The set of all the hidden variables is denoted by Z. In this case the likelihood function is expressed as:

$$\hat{\Omega} = \max_{\Omega} \log \mathcal{L}(\Omega \mid \mathsf{X}, \mathsf{Z}) \ , \tag{2.18}$$

where $\mathcal{L}(\Omega \mid \mathsf{X}, \mathsf{Z}) = p(\mathsf{X}, \mathsf{Z} \mid \Omega)$. Because we are given only the incomplete dataset X, the knowledge available for the hidden variables Z is inferred from the posteriori distribution $p(\mathsf{Z} \mid \mathsf{X}, \Omega)$. The expected value is found by evaluating the posteriori distribution using the previous (or initial) values of the parameters, denoted by Ω^{old}. Then, this posteriori distribution is used to calculate the complete likelihood function evaluated for some general parameter value Ω (see equation (2.19)). This process corresponds to the Expectation step.

$$Q(\Omega, \Omega^{\text{old}}) = \sum_{\mathsf{Z}} p(\mathsf{Z} \mid \mathsf{X}, \Omega) \log \mathcal{L}(\Omega \mid \mathsf{X}, \mathsf{Z}) \ , \tag{2.19}$$

Then, the subsequent M step, uses the result of the E step to evaluate $\Omega^{\text{new}} = \max_{\Omega} Q(\Omega, \Omega^{\text{old}})$. The process is repeated iteratively until a convergence criterion is satisfied.

2.4.2 Metrics for Comparing Histograms

Image retrieval systems usually search for the optimal descriptors that allow to perform a specific task. In the case of exploring large image databases, one of the main objectives is to find similar images. This goal can be easily reached if we are able to compare two images in a quantitative manner. The question is how to measure the resemblance between images. Of course, the descriptors that represent images are "optimal" depending on the application and the characteristics to compare. Geometric information is often discarded, since color histograms usually encode enough information to find similar images based on their appearance.

A color histogram is a mapping from a set of D-dimensional vectors, usually representing a color with 3 components (RGB, La^*b^*, HSV) to an N-dimensional vector. This

vector fixes some partitions, called bins, on the D-dimensional space from previously accorded boundaries. Each bin stores a measure of the number of colors that fall into the specific partition. For example, if the pixels of a color image are treated in the HSV color space and we want to obtain the color histogram, we must define the number of partitions for each channel. Let us say we choose 5 partitions for the hue channel, 5 for the saturation channel and 3 for the value channel. Then, the result is a 75-vector representing an estimate of the probability distribution of the colors. Thus, a histogram P is denoted by $P = \{(p_1, w_{p_1}), \cdots, (p_N, wp_N)\}$, where the N-dimensional vector $p = [p_1, \cdots, p_N]$ contains the bin centers and $w = [w_1, \cdots, w_N]$ stores the number of pixels falling in each partition. The works of Rubner *et al.* [Rubner 2000] and Muselet [Muselet 2004] present a good compilation of the techniques used for histogram comparison. In order to justify the choice we made for the algorithm presented in chapter 3, we summarize some of the similarity measures commonly used and their advantages and disadvantages in the present context.

The similarity measures can be classified in two main categories: those that relate the corresponding bins of two histograms, formally known as *bin–to–bin* measures and the metrics that compare histograms based on a ground distance between the bins of different histograms. These are called *cross–bin* measures.

Bin–to–bin measures. A set of popular metrics used to compare discrete distributions is formulated by the *Minkowski distance*. Given two histograms $P = \{(p_1, w_{p_1}), \cdots, (p_N, w_{p_N})\}$ and $Q = \{(q_1, w_{q_1}), \cdots, (q_N, w_{q_N})\}$ this metric is defined as follows:

$$L_K(P, Q) = \sqrt[K]{\left(\sum_{i=1}^{N} |w_{p_i} - w_{q_i}|^K \right)} \quad . \tag{2.20}$$

When $K = 1$, the measure L_1 is called the Manhattan distance, while with $K = 2$, it becomes the classical Euclidean distance. The L_∞ distance is also commonly used. Among the advantages of this technique of comparison is clearly its easy implementation and comprehension. The fact that the measure corresponds to a true metric (in the formal mathematical meaning of the word, *i.e.* it meets the non-negativity, symmetry, subadditivity and identity of indiscernible properties), makes of this measure one of the favorites for matching simple distributions.

Another distance proposed for histogram comparison is the *Histogram Intersection*:

$$I(P, Q) = 1 - \frac{\sum_i \min(w_{p_i}, w_{q_i})}{\sum_i w_{q_i}} \quad . \tag{2.21}$$

Although this distance is not a metric (it does not comply to the symmetry property), it is well known for its ability to handle partial matches when the areas of P and Q are different. The *Kullback–Leiber divergence* is another measure used to compare probability distributions. It is derived from information theory concepts. It is described for the continuous case in

section 3.1.3 . The KL divergence between two histograms ($\mathsf{KL}(P, Q)$) does not correspond to a true metric, but it is, intuitively, a measure of how inefficient it would be to code the histogram P using a code based on histogram Q rather than using a code based on P.

The problem with the bin–to–bin distances is that, in general, they are extremely sensitive to the bin position and size. For example, in his doctoral dissertation, Muselet [Muselet 2005] shows that images of the same object illuminated under different light conditions could exhibit histograms with similar shapes but shifted by some bins. If using this kind of measures in those cases, images of the same object could present a poor resemblance.

Cross–bin measures. Some measures include in their computations a *ground distance* that takes into account changes relative to different bins of the histograms. Information about the position of the bins is also encoded, besides the number of counts falling into a bin. For example the *Quadratic–form* distance, also called the *Mahalanobis* distance, connects the different bins of two histograms through the covariance matrix or the similarity matrix (S). The latter is dependent on the ground distance function. The Quadratic–form distance is defined by:

$$D(P, Q) = \sqrt{(\mathbf{w}_p - \mathbf{w}_q)^{\mathsf{T}} S^{-1} (\mathbf{w}_p - \mathbf{w}_q)} \ , \tag{2.22}$$

for $\mathbf{w}_p = [w_{p_1}, \cdots, w_{p_N}]$ and $\mathbf{w}_q = [w_{q_1}, \cdots, w_{q_N}]$. We can also find in this category the Earth Mover's Distance (EMD). It is one of the most appreciated similarity measures in image retrieval challenges given that it is almost invariant to histogram shifting. Fast implementations suitable for image search in large database are proposed by [Pele 2009]. This distance is discussed more extensively in section 3.1.3.

CHAPTER 3

Classification of Images based on the Sky Appearance

Contents

Résumé. Dans les systèmes de vision par ordinateur qui opèrent dans des milieux extérieurs, la compréhension de la nature de l'illumination est un facteur déterminant. Par exemple, des applications pratiques comme la navigation intelligente de véhicules, la détection automatique des objets, la reconnaissance de visages, etc., requiérent la plus part du temps une estimation approximée ou des hypothèses sur la source d'illumination dans la scène. Dans le cas particulier des scènes d'extérieur, les rayons émis par le soleil se reflètent et se réfractent dans l'atmosphère, créant ainsi une illumination globale qui détermine la façon de percevoir les objets qui les compose. En Vision par Ordinateur, l'exploitation du ciel comme la principale source d'illumination est étudiée depuis de nombreuses années. Certaines des premières solutions proposées reposent sur des approximations utilisant la couleur. Les autres utilisent des modèles physiques, toutefois ils admettent que les paramètres photométriques et géométriques de l'appareil photo soient les mêmes ou du moins soient connus. Dans ce chapitre, nous proposons une méthode simple et efficace permettant de grouper les images ayant des conditions d'illumination similaires. On propose deux algorithmes fondés sur des outils statistiques afin de comparer les résultats obtenus par les deux procédures. Tout d'abord nous obtenons une segmentation de la région de l'image correspondant au ciel. Ensuite, nous estimons les paramètres des modèles statistiques utilisés pour représenter les pixels du ciel. Finalement, nous calculons les similarités entre les modèles pour trouver des apparences similaires.

Understanding natural illumination in outdoor images has acquired a critical importance for computer vision systems. Applications such as self-driven cars, image composing, automatic object detection or face recognition need most of the time a description of the illumination to successfully perform in real world conditions. In the case of outdoor scenes, when looking at an image, sky appearance encodes most of the information relative to the illumination of the scene. In fact, trained personnel such as plane pilots, sailors or meteorologists are able to deduce important information about the weather conditions or the time of day just glancing up to the sky. Less skilled people can say if it is night or day, sunny or cloudy, but as human beings the omnipresent sky gives us always important clues about the nature of the illumination. Considering this remarkable influence of the sky appearance on the lighting of the scene, we introduce in this chapter a system for automatic classification of outdoor images using information from visible parts of the sky and minimum information about the cameras.

Frequently, photos of outdoor scenes show in the background plane portions of the sky. These regions give us important clues to extract information about the illumination of the scene. The solar beams interact with the atmosphere creating an ambient lighting that determines the appearance of the objects. Our objective consists in grouping images with similar sky appearance. This might be useful for applications like automatic photomontage (see figure 3), augmented reality, or efficient image searching. To do this, we propose two methods based on statistical tools. Both of them can be described by a 3-stages pipeline. First we segment the sky region from the image. Then we find the parameters of statistical models for the colored–pixels of the sky. Finally, we compare the models using an appropriate similarity measure. We shall explain in detail each of these stages in the next section, then we shall present comparisons of both methods and their results, along with the corresponding discussion.

3.1 Statistical Representation of the Sky Appearance

Different alternatives to represent the natural illumination of outdoor scenes have been proposed by researchers in diverse areas. Following the discussion presented in Lalonde's doctoral dissertation [Lalonde 2011], where the approaches are classified as physically–based representations, environment maps and statistical representations (*cf.* 2.2), the models proposed in this work can be fitted in the last–mentioned group. Indeed, we represent the pixels from portions of visible sky by probability distributions in the appropriate color space. We describe our approach as a process involving three fundamental steps: segmentation, modeling and classification. Each of these parts is explored in the next paragraphs.

Figure 3.1: Example of photomontages where illumination conditions between the original image and the backgound of the query images are different (top right) and when the sky appearance coincides (bottom right). Which photo montage is more believable?

3.1.1 Sky Segmentation

Our goal is to compare images based on the appearance of the sky. Consequently, the first step in our approach tries to isolate this region of interest on the image. To find this sky region, we used the method proposed by [Hoiem 2005]. In their work authors introduce a method that allows us to decompose any outdoor image in three main components: the sky, the ground and the objects placed on the surface of the ground. The main contribution of this algorithm is to segment the images into coherent regions, using an *a priori* about the geometric distribution of the scene.

The mentioned approach uses multiple features present in outdoor scenes, including color, texture, location and shape feature descriptors. It also uses a coarse description of the 3D structure of the scene. The first step to estimate the region of interest (ROI), in our case the sky, consists in localizing contiguous zones of uniform color. This oversegmentation creates uniform regions called superpixels [Ren 2003]. For a given number of regions, the algorithm searches to maximize the joint likelihood that homogenizes all the regions at the same time. This operation is performed using a greedy search algorithm: authors propose to assign weights obtained from an affinity function to relate pairs of randomly selected superpixels. The procedure is repeated iteratively until a solution is found.

The success of the approach resides in the computation of the affinity function. This estimation is done using a supervised learning approach. Thousands of previously labeled–images are used to calculate the codomain of the affinity function. This prior information is used to train a classifier of the type Adaboost [Freund 1995]. Same–labeled and different–labeled superpixels are arranged to form the training data set. Based on the absolute differences given by the features, the classifier learns a measure of similarity between two superpixels. In this way, larger regions are composed, given as result a segmentation algorithm robust against common sources of error as for example, changing illumination, disparity on the camera colors, occlusions, etc. Figure 3.2 shows some results of the sky segmentation algorithm applied on images from our databases.

3.1.2 Modeling the Sky Pixels

Once the sky pixels are extracted from the image, they are initially stored as intensity values in the RGB color space. However, this representation does not seem to be adequate to analyse this kind of variables. Most of the values fall into a small region of the RGB space (see figure 3.3). The choice of a more discriminative color space plays an important role in our method. The CIE (Commission Internationale de l'Éclairage) [C.T 1994] has created several standards and nowadays, the last released model (CIECAM02) has reached a high efficiency performance in color–related applications as well as a high degree of complexity [C.T 2002]. This color appearance model allows a trustworthy representation of the real world phenomena. However the parameters describing the model are extremely difficult to acquire, sometimes requiring a full detailed knowledge of the conditions during the acquisition. As alter-

Figure 3.2: Sky segmentation on some images of the database.

native, the CIE La^*b^* space and the xyY space used by [Lalonde 2007] and [Lalonde 2008] are simpler (but also effective) models derived from previous CIE's works. Although these color representations are used in a vast number of works, it is important to take precautions when they are applied. For example, the color appearance models above mentioned are always attached to a predetermined "reference white", usually unknown. In these cases the color systems are dependent on the imaging conditions [Gevers 2001]. This characteristic is alleviated by the fact that illumination of outdoor scenes came usually from the same source (the sky). We also establish the first assumption for our algorithm: A daylight illuminant is present in all photos. The validity of this hypothesis is supported by a previous filtering–step of the image collection, where the images taken at night and those with artificial lighting are manually removed. To process the pixels, we chose the CIE La^*b^* color space because of its known ability to separate the brightness from the chrominance information and because it is closely related to the way how humans perceive color. The color space conversion is done under the assumption previously mentioned. The "reference white" used in the color space transformation is the same for all images (Illuminant D65).

To model the sky pixels we use two alternatives representations. The first one consists of discrete probability distributions, or 3D histograms and the second uses a mixture of Gaussian distributions fitted to the sky pixels. We describe these two choices in the next paragraphs.

3D Histograms. For each image, a discrete probability distribution is built from the sky pixels. This representation is visualized as a histogram of variable bin size in a 3D space. The La^*b^* space is quantized using 4 bins for the luminance channel and 8 bins for each chrominance channel. This is explained because changes on brightness play a role less de-

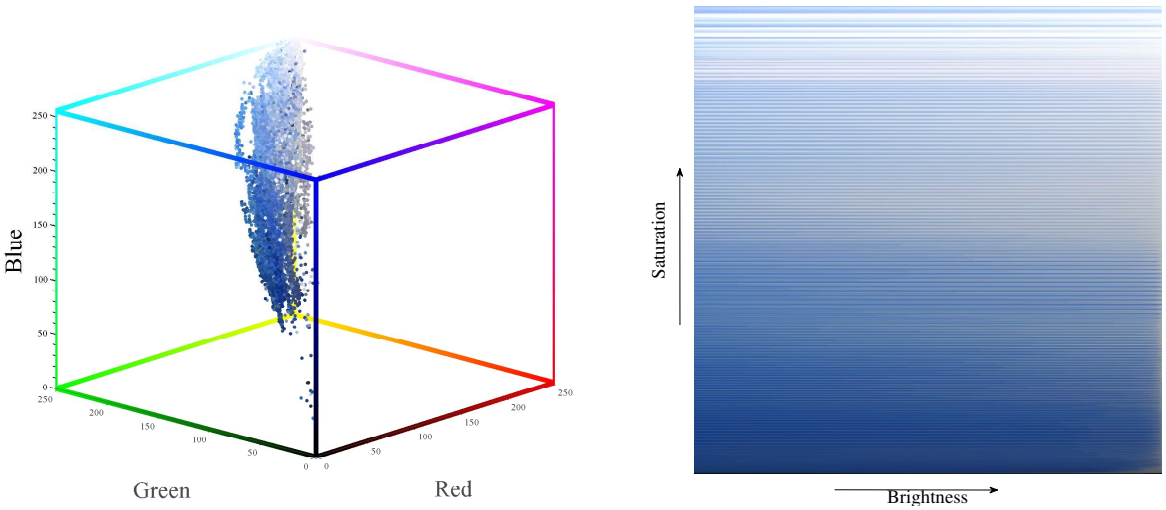


Figure 3.3: Pixels from the sky extracted randomly from 1000 outdoor images. Their distribution in the RGB cube is shown on the left and the pixels ranged by their bightness and saturation are represented on the right.

terminant than color variations during our classification. In consequence, there are 4×8^2 quantized values (bins) for the sky colors. Due to the high concentration of some values over small regions of the possible range of luminance and chrominance, the width of the bins is not uniform, and it is calculated using a training set: we extract pixels corresponding to the sky from 1000 images and we select the ranges that generate a uniform distribution: *i.e.* the ranges where the number of elements in all the bins are approximately equal, for all bins. The histograms are normalized. Figure 3.4 shows some typical histograms. One can observe that the distributions of sky with uniform color exhibit one or two main modes while cloudy skies show multiple modes in their histograms.

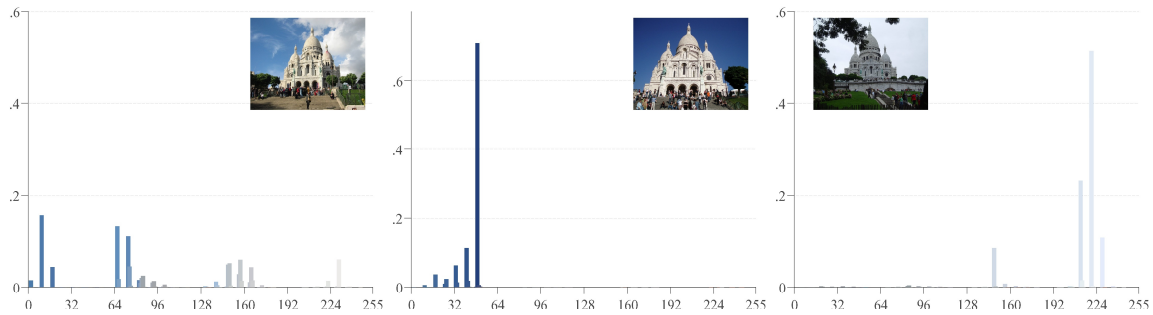


Figure 3.4: Representation of the color histograms.

Gaussian mixture models (GGM). According to our observations, the sky pixels in each image often show multi-modal distributions. They are modeled by fitting a mixture of Gaussian distributions to the data. Each set of pixels from the sky is represented by an $N \times 3$ matrix denoted X. It stores the La^*b^* values for the sky's pixels. Following the description of the multiple Gaussian distributions given in section 2.4.1, the model is composed by K Gaussian densities:

$$M(\mathsf{X}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x} \mid \mu_k, \Sigma_k\right) \quad , \tag{3.1}$$

In our implementation $K \in \{2, 3, 4\}$. This empirically-found set of values is based on the observation of the pixels in the La^*b^* color space and the fact that the maximum number of modes in the histograms is never higher than 4. Each Gaussian distribution belonging to the linear combination of K distributions is parameterized by its mean $\mu_{(1\times3)}$, its covariance matrix $\Sigma_{(3\times3)}$ and the mixing coefficient π. Parameters in equation (3.1) are found using the well–known algorithm of Expectation–Maximization (EM) [Bishop 2006] (see section 2.4.1. The output of this method corresponds to the variables π_k, μ_k, Σ_k that describe the model in such a way that the likelihood of the solution given the data is maximized. Our model is formulated in terms of a joint probability distribution of the sky pixels X, some hidden (and unknown) variables Z and the goal is to find the set of values $\theta = \{\pi_k, \mu_k, \Sigma_k\}$ that maximize the likelihood of the function (*cf.* 2.4.1). The latent variable matrix $\mathsf{Z}_{(N\times K)}$ contains for each row a K-dimensional binary vector representing the "state" for a particular sample respect to the variables.

The EM algorithm applied to the sky pixels X can be summarized this way:

1. Select random initial values for π_k, μ_k and Σ_k.

2. **Step E**. Calculate the joint probability distribution of the latent variables, given the sky colors and the initial values, $P(\mathsf{Z} \mid \mathsf{X}, \theta^{\text{last}})$.

$$P(\mathsf{Z}|\mathsf{X}, \theta^{\text{last}}) = \frac{\pi_k \mathcal{N}\left(\mathsf{X}_n \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathsf{X}_n \mid \mu_j, \Sigma_j\right)} \quad .$$

3. **Step M**. Find the new mixing coefficients π_k, centroids μ_k, and correlation matrix Σ_k that define each Gaussian.

4. Run alternate steps 2 and 3 until convergence (e.g. until no large changes in π_k, μ_k and Σ_k occur).

The number of Gaussian distributions (K) used by the mixture, *i.e.* the dimension of the model, is determined by using the Akaike information criterion (AIC) [Akaike 1974, Bishop 2006]. This measure determines a good trade–off between the relative model's goodness of fit to the data and the number of adjustable parameters. We compare the AIC measures for different values of K in equation (3.1), the model with best performance is chosen.

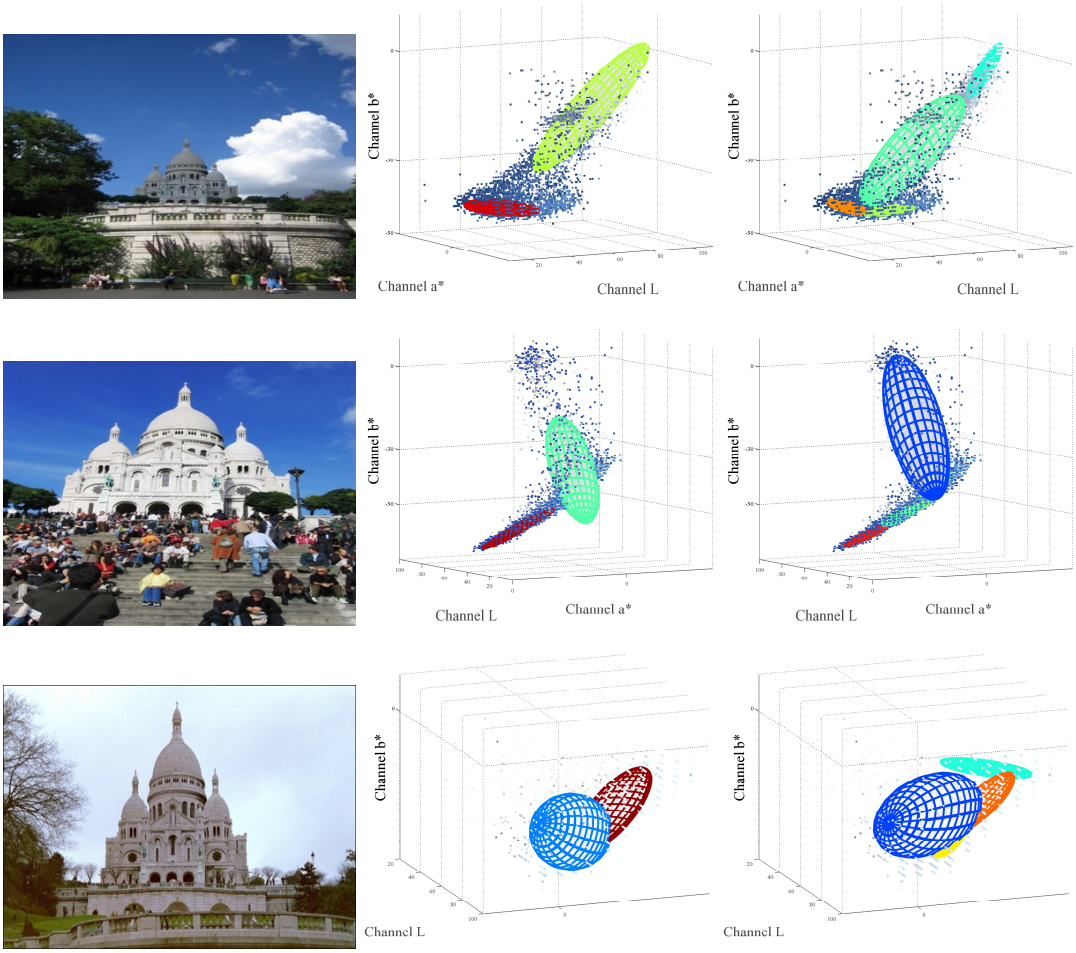


Figure 3.5: Mixture of Gaussian distributions fitted to the La^*b^* pixels. Colors on the ellipsoids that represent the probability distributions are proportional to the mixture coefficient of Gaussian distributions (hot colors mean higher weight while cold colors mean low values for the mixing coefficients).

3.1.3 Comparison of images based on sky appearance

Once the sky model for each photo is estimated, we proceed to compare different images. Depending on the representation (color histogram or GMM), we use two different metrics. Even if both models are probability distributions, for the case of the color histograms we use a dissimilarity measure according to this discrete representation, the Earth Mover's Distance (EMD). For the case of mixture of Gaussian distributions we compare the continuous distributions using a measure very common in the Information Theory community, the Kullback–Leibler divergence.

Earth–mover's distance (EMD). This similarity measure proposed first by [Rubner 1998] is widely used in object retrieval problems to improve the performance of classifiers usually based on color and texture. A more efficient reformulation of the EMD using the L_1 metric was proposed later by Ling and Okada [Ling 2007]. The Earth Mover's Distance defines a consistent measure of similarity between two distributions in a space with a given *ground distance*. It addresses several difficulties found in histogram comparison, for example, when using bin to bin metrics (*i.e.* all the variations of the Minkowski distance, including the Euclidean distance, *cf.* section 2.4.2). In those cases the traditional metrics do not take into account histogram shifts even though often the forms are similar; this behaviour is common when two histograms correspond to images of the same object taken with different illumination conditions or different exposures. To overcome these difficulties, the EMD proposes a solution based on the well–known *transportation problem*. The objective of this problem is to minimize the distance traveled by an undetermined group of providers supplying different points with limited storage capacity. Bringing this concept to our framework, the intuitive idea behind the EMD algorithm is to find the minimum amount of "work" required to transform a 3D histogram into another. Formally, given two histograms with n bins ($P = \{(p_1, w_{p_1}), \ldots, (p_n, w_{p_n})\}$ and $Q = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\}$) , each of them characterized by the bin centers (p_j, q_j) and the amount of pixels falling in the j^{th} bin (w_{p_j}, w_{q_j}), the matrix $\mathsf{D} = [d_{ij}]$ is a symmetric matrix containing information on the position of the bins, and the cost of "moving" pixels from the bin i to the bin j (for example, the distance of moving a pixel from one to a consecutive bin is 1) . The goal consists in finding a flow $\mathsf{F} = [f_{ij}]$, where f_{ij} is the flow between p_i and q_j that minimizes the overall cost:

$$\text{work}(P, Q, \mathsf{F}) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} \quad , \tag{3.2}$$

under the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{3.3}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}, \qquad 1 \leq i \leq m \tag{3.4}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j}, \qquad 1 \leq j \leq n \tag{3.5}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \tag{3.6}$$

Once the flow matrix F is found the EMD is defined by:

$$\mathrm{EMD}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{3.7}$$

To solve the linear optimization problem authors in [Rubner 1998] proposed a linear programming solution called the Transportation Simplex algorithm. This solution consists of a modified form of the simplex algorithm which reduces the complexity of the operations needed to maintain the flow matrix, by taking advantage of its structure. A more optimum version of the algorithm was proposed in [Ling 2007], exploiting the use of the L_1 metric as *ground distance* when possible. For the case of histograms this simplification is very convenient, since the processing time is largly reduced.

Kullback–Leibler divergence. To measure the difference between two or more mixture of Gaussian distributions, in our context, the Kullback–Leibler divergence (KL) might be a good option, although it does not possess the property of symmetry. Given two probability distributions $p(x)$ and $q(x)$ the KL divergence is defined by:

$$\mathsf{KL}(p||q) = -\int p(x) ln\left\{\frac{q(x)}{p(x)}\right\} dx. \tag{3.8}$$

It is widely proven that when the distributions are formed by a single Gaussian, equation (3.8) can be expressed in closed-form. However, in the case of a Gaussian mixture, it is difficult to find an analytically tractable expression or even more, a computer algorithm to solve this problem efficiently. Therefore the alternative is to use a good approximation. The work of Hershey and Olsen [Hershey 2007] offers a complete revision of the subject. The authors also make a benchmarking test with the most popular algorithms and propose a novel estimation. Up to date, the only method for estimating $\mathsf{KL}(p||q)$ with arbitrary accuracy when $p(x)$ and $q(x)$ correspond to mixtures of Gaussian distributions, is the Monte Carlo simulation. Nevertheless, other approximations may be valid, depending on the context. For

example, a commonly used approximation is the simplification of the Gaussian mixtures $p(x)$ and $q(x)$ by a single Gaussian $\tilde{p}(x)$ and $\tilde{q}(x)$. The mean and covariance estimated are:

$$\begin{aligned}\mu_{\tilde{p}} &= \sum\nolimits_a \pi_a \mu_a \\ \Sigma_{\tilde{p}} &= \sum\nolimits_a \pi_a \left(\Sigma_a + (\mu_a - \mu_{\tilde{p}})(\mu_a - \mu_{\tilde{p}})^{\mathsf{T}}\right).\end{aligned} \tag{3.9}$$

In this case, the KL divergence (KL_{sim}) is calculated using the estimated mean ($\mu_{\tilde{p}}$) and variance ($\Sigma_{\tilde{p}}$). Hershey and Olsen use variational methods to find a better approximation of the KL divergence. One of their contributions is a measure that satisfies the symmetry property but not the property of positivity. In this case, the approximated divergence $\mathsf{KL}_{\text{app}}(p||q)$ is given by:

$$\mathsf{KL}_{\text{app}}(p||q) = \sum_a \pi_a \log \frac{\sum'_a \pi'_a e^{-\mathsf{KL}(p||p')}}{\sum_b \omega_b e^{-\mathsf{KL}(p||q)}}. \tag{3.10}$$

This value could be seen as a measure of similarity between two distributions. Hershey and Olsen's contribution allows us to compare two sky-gaussian models keeping intact all the characteristics of the Gaussians that compose the mixtures.

3.2 Experiments and Results

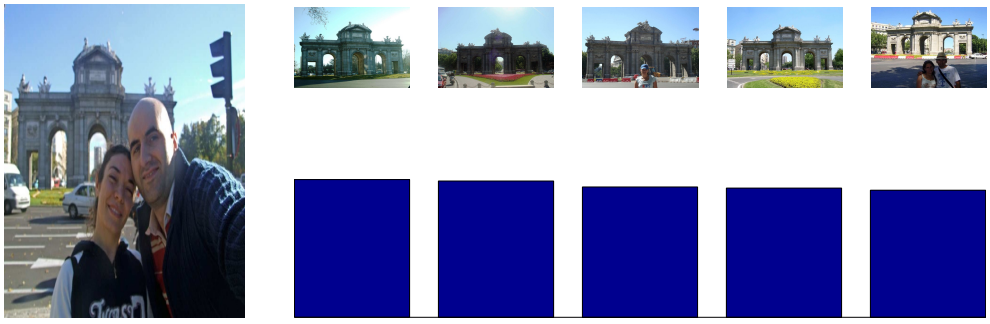
This section presents the experiments carried out to validate our approaches and it also shows the results of the estimated models and the comparison metrics previously described. To test these methods, we compiled three databases. Each database consists of hundreds of outdoor images from a particular location downloaded from the Flickr website[1]. The downloaded images were originally shared by their owners under a Creative Commons license (CC) which guarantees their use for research purposes. The first database is composed of 1250 images of the *Sacré Cœur's* Cathedral in Paris (France), the second one consists of 2120 images of the tower of Pisa in Italy. The third one is composed of 750 images of the *Puerta de Alcalá* in Madrid (Spain). All images were downloaded automatically by using the script available from [Hays 2007]. When available, the metadata, *i.e.* the EXIF tags, were kept. Images that do not correspond to the desirable scenes, taken during the night or using artificial lighting were manually removed.

Our goal is to compare two or more outdoor images using the models before described: the histograms and the EMD distance for *method 1* and the Gaussian mixture model using the KL divergence for *method 2*. In order to establish a ground truth, we classified all the images from each database by hand according to three previously defined tags: *sunny images* (*si*), *partially cloudy images* (*pci*) and *completely cloudy images* (*cci*). To prevent the influence of subjective criteria, we define a sunny image as a photograph where the sky is completely visible without interference of clouds. A partially cloudy image represents an image in which

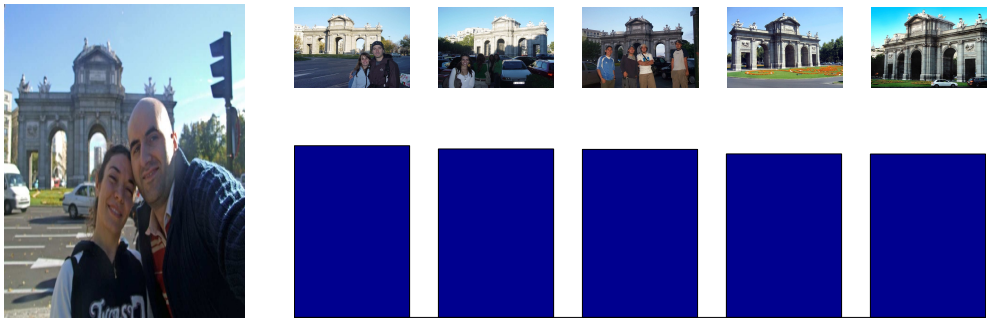
[1] http://www.flickr.com/

some clouds are visible, or some fog is present. If the entire surface of the sky projected on the image shows a well defined range of colors between white and gray, we say that it corresponds to the completely cloudy image group. Despite the use of these general policies to perform the manually classification, it is important to emphasize that the subjective nature of this task remains.

Our first experiment consists in comparing the distances between one randomly–selected query image and all the other images present on the same database. When using method 2, low values for the KL divergence represent high similarity between the models while high values imply poorly correlated distributions. For the case of method 1, as well, the low magnitude of the EMD implies high similarity and vice versa. Figure 3.6, 3.7 and 3.8 shows a query image along with the most similar images found on our databases. The query is done using the GM model and the KL_{app} divergence for figures 3.6(a), 3.7(a), 3.8(a) and 3D histograms with EMD for figures 3.6(b), 3.7(b), 3.8(b). Note that values of the bars are proportional to the inverse of normalized value of the measure. Normalization is done between zero and one: the maximum value (1) is the measure of the image with itself. Additional results showing more extensive searches are shown in Appendix A.



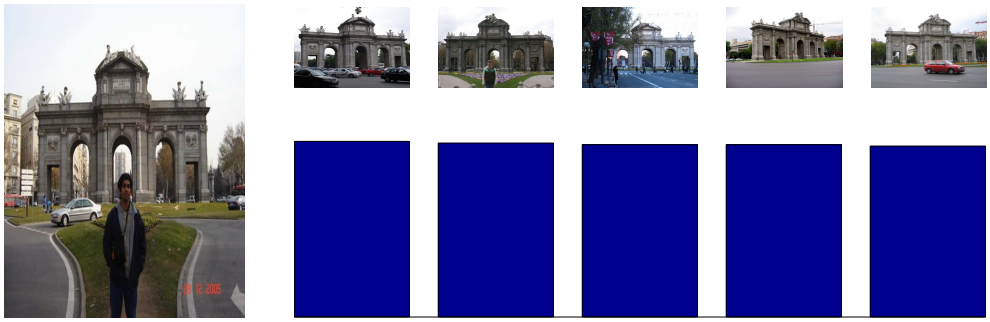
(a)



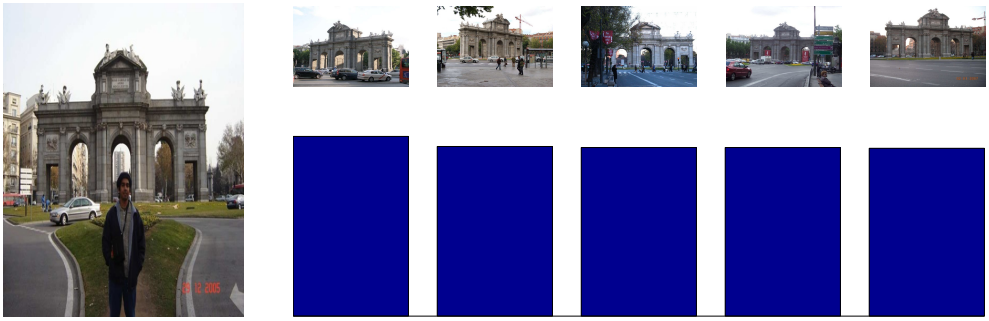
(b)

Figure 3.6: Query image and the six closest images using: method 1 for (3.6(a) and method 2 for 3.6(b). Bar sizes represent the normalized metric.

A second experiment comparing all the images belonging to one particular group to the others within the same and different groups was done. These groups correspond to the manu-



(a)



(b)

Figure 3.7: *cf.* figure 3.6.

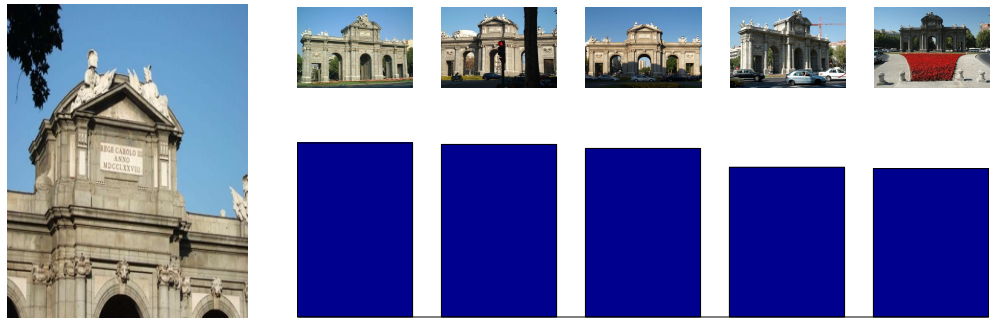
ally classification previously described. Given two probability distributions $p_K(x)$ and $q_K(x)$ tagged with the index $K = \{\text{si}, \text{cci}, \text{pci}\}$, we define respectively the intra–class distance and the inter-class distance as:

$$d_{\text{inter}} = D(p_l(x), q_m(x)) \qquad \text{for } l = m \tag{3.11}$$

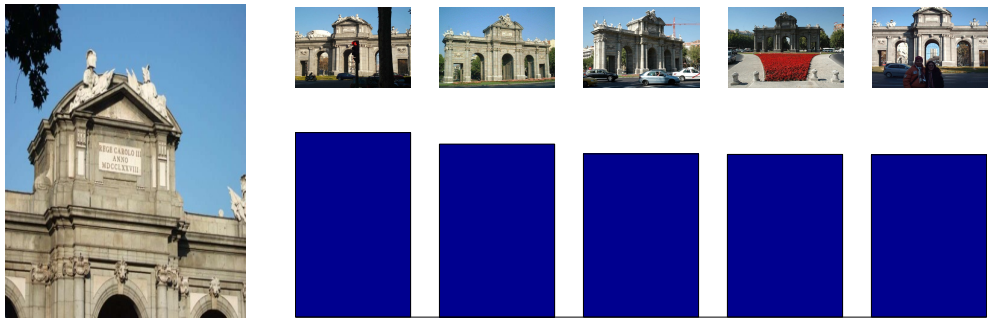
$$d_{\text{intra}} = D(p_l(x), q_m(x)) \qquad \text{for } l \neq m \ , \tag{3.12}$$

where $D(p(x), q(x))$ can be either the EMD between two histograms or the KL divergence for Gaussian mixture models.

One query image was compared to the other ones belonging to the same group (intra–class distance) and those attached to other groups (inter-class distance). Figure 3.9 shows the intra–class and inter–class mean distances computed for 170 images of the group *si* using the $\text{KL}_{\text{simple}}$ divergence (fig. 3.9(a)), the KL_{app} divergence (fig. 3.9(b)) and the EMD (fig. 3.9(c)). For this measure, the images belong to the *Sacre Coeur's* database. The experiment reveals us that high similarity is found when comparing images from the group *si* with photographs of the same class. In all the three cases the comparison using intra–class distance is, in most of the cases, the lowest (high similarity). The inter–class distance between the group *si* relative to the groups *pci* and *cci* is very close between them, but still separable and distinguishable. The best performance is reached when using the KL_{app} divergence, result are understandable given that this measure takes into account more complex and detailed mod-



(a)



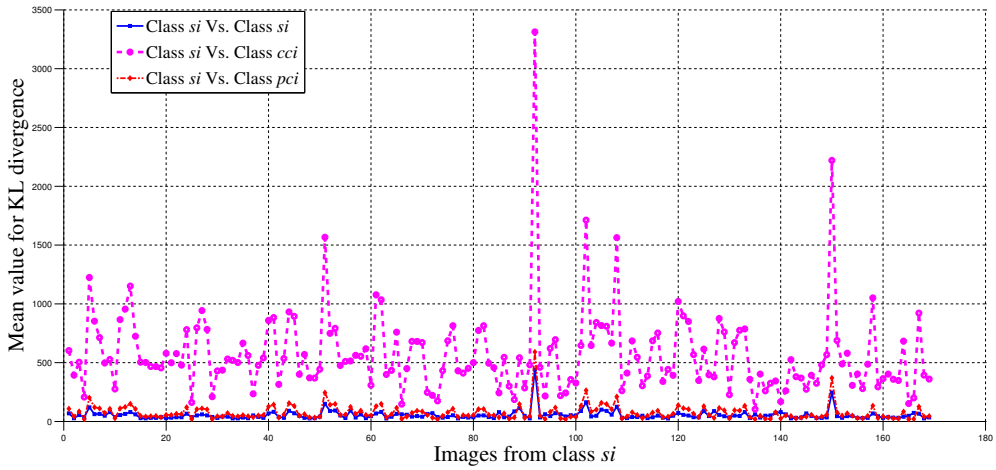
(b)

Figure 3.8: *cf.* figure 3.6.

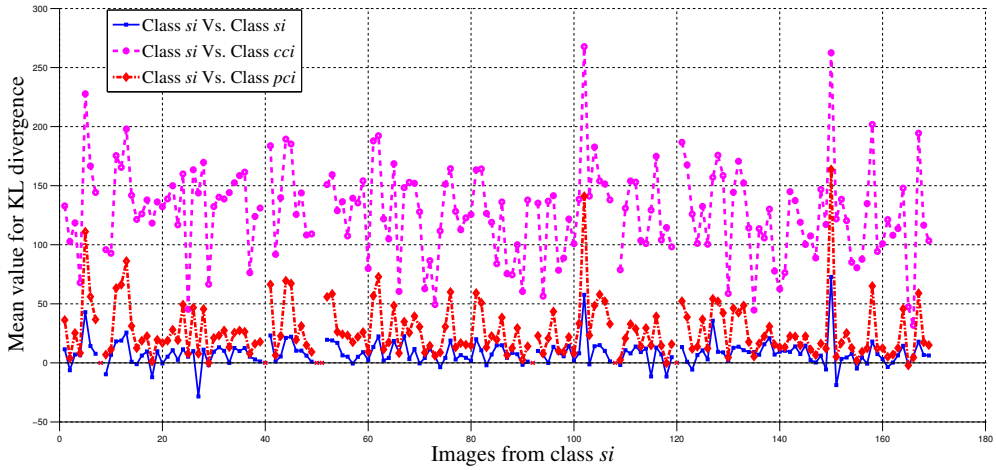
els. The KL divergence and the model using Gaussian distributions (method 2) remain as a better option than the 3D histograms using the EMD (method 1). This result can be explained because in method 2 we use the Akaike information criterion to establish automatically the number of parameters composing the mixture of Gaussians model, while in method 1 the number of bins for all the histograms is a constant for all the images of the collection. Thus, finer details can be represented using method 2. The two first plots in figure 3.9 show a large gap between the intra-class distance of *si* and *cci*, which results on a better discrimination property for the system.

Table 3.2 presents the average of the inter and intra-class distances for all the previously tagged groups. Columns correspond to the average of the inter or intra–class distance using the metrics $\mathsf{KL}_{\text{simple}}$, KL_{app} and EMD, for one of the databases. It is evident that low scores are achieved for intra–class distances, and the discrimination capacity is improved using method 2 (KL divergence along with Gaussian models).

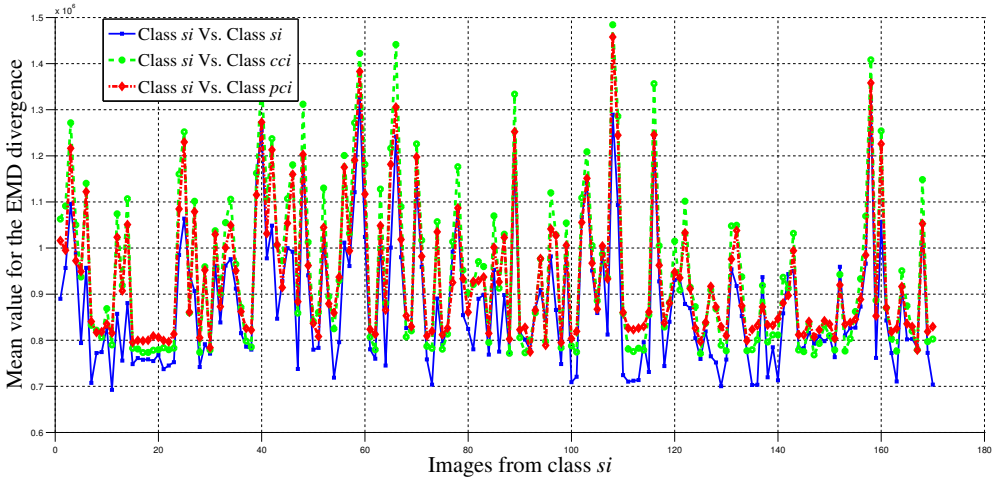
We have to remark that our algorithm failed to correctly classify a few images when using KL_{app}. Infinite values were assigned sporadically. This is a consequence of the convergence of the approximation formula towards infinity. When the exponents of the exponential function are of great value the function tends to zero. This result is problematic for the implementation, as a consequence these results have been ignored in the final results.



(a)



(b)



(c)

Figure 3.9: Mean values for the intra-class and inter-class divergence, (a) using the $\mathsf{KL}_{\text{simple}}$ divergence and (b) using the KL_{app} divergence and (c) using EMD on the histograms.

	GMM + $\mathsf{KL}_{\text{simple}}$			GMM + KL_{app}			Hist. + EMD ($\times 10^6$)		
	si	*pci*	*cci*	*si*	*pci*	*cci*	*si*	*pci*	*cci*
si	45.60	50.95	351.63	4.27	13.73	97.87	0.88	0.95	1.19
pci	55.25	38.76	201.76	15.29	3.95	38.26	1.42	1.3	1.56
cci	222.11	63.72	46.33	75.32	22.39	4.55	1.55	1.23	0.77

Table 3.1: Average of the inter and intra-class distance for database 1 *Sacre Coeur*. For each image belonging to any of the groups listed on the rows *si, pci, cci*, we compare it to all the other images attached to the same group (intra–class) or different groups (inter–class). The average over all the pairs is shown.

3.3 Discussion

In this chapter we propose simple and effective algorithms for grouping outdoor images with similar sky appearance. The approaches do not uses information about the camera parameters or about their position with respect to the scene and they work with images collected from the Internet. The proposed methods are based on a 3-steps pipeline: sky segmentation, modeling sky regions and a robust comparison between these models. To estimate the model, we formulate two alternative representations: color histograms and mixture of Gaussian distributions, both applied on the sky pixels transformed on the La^*b^* color space. Comparison of models is made using specific similarity measures: the Earth Mover's Distance (EMD) and the KL divergence.

The approach described by method 1 is inspired by algorithms specifically designed for image query in large databases. Despite its good performance, we observed during its implementation that it is highly dependent on the quantization used to create the bins of the histograms. (The quantization of the samples space in $4 \times 8 \times 8$ bins was the one that perform the best results, but it was empirically found). Also the boundaries for the bins were found by a data-driven method which requires a preliminary step. One could eventually use a criterion to establish the right number of bins and ranges for the histogram but its computation might become complex. In this sense, the proposed criterion should act similar than the AIC used in the second algorithm. On the other hand, we present a second approach, *method 2*, using a GMM and the KL divergence. For this algorithm we formulate two derivations: using a simple approximation of the KL divergence and a more elaborated approximation of this function.

Since both methods generate good quality results, we can compare them putting side by side each of the steps involved in the process: While method 1 uses histograms for modeling sky pixels, method 2 finds mixture of Gaussian distributions that fit the data. Histograms suffer of the problem previously described (how to select the right number of bins to describe a set of data) and, in our case this inconvenient is worked out by a pre-processing step. In the case of method 2, we do not perform this offline process and we can fit directly the sky pixels to the GMM. The new difficulty is that fitting Gaussian mixture models using the EM algo-

rithm is not an extremely fast process (depending on the number of Gaussian distributions in the model and on the number of samples). In our case, for samples of around 500000 pixels it took approximately fifteen seconds to fit a GMM of 4 components. It does not seem to be a large processing time, but when using large databases it could become intractable. However Gaussian mixtures for the sky pixels seems to describe better the behaviour of images, particularly when the sky presents uniform colors. That happens because histograms tend to lose information, specially when the data is stored in only a few bins. On the comparison step, the EMD has been proven to be a good measure of similarity, however its performance its directly linked to the modeling step (wrong descriptors lead to bad comparisons). On the side of the KL divergence the derivation using the KL_{app} presents better results as expected (it uses all the information available in the mixture of Gaussians).

The empirical approaches discussed in this chapter show a good performance but it is clear that without more information about the scene it is difficult to infer accurate knowledge of the illumination conditions and in general of the photometric properties of the scene. Inspired by the geometric 3D reconstruction process, we propose to exploit the 3D information available on the scene, given that multiple views of the same object are available, to infer more precise and accurate photometric information. The next two chapters describe theses processes.

CHAPTER 4

Estimation of the Camera Response Function

Contents

Résumé. En explorant des nouvelles façons d'interpréter l'illumination dans des ambiances extérieurs nous avons présenté dans le chapitre précédent une méthode pour classifier les images en utilisant les régions visibles du ciel. Dans ce cas, nous avons exposé l'importance de l'illumination globale pour la compréhension de la scène. Malgré les bons résultats de cette approche exploratrice, nous utilisons seulement une partie de l'information disponible dans les images, c'est à dire, l'information purement photométrique. Dans ce chapitre, nous proposons d'incorporer d'autres composants relatifs à la formation de l'image afin d'exploiter une collection d'images. Étant donné que nous avons accès à une large base d'images qui correspondent à la même scène, nous reconstruisons la géométrie 3D pour retracer avec un certain dégrée de détail le processus de formation de l'image. En général, ce processus est décomposé par trois composants qui interagissent entre eux pour générer l'image. Ces parties sont : l'illumination, les propriétés de réflectance de l'objet et les propriétés radiométriques de l'appareil photo. Tout d'abord nous allons concentrer nos analyses sur ce dernier aspect, en conservant aussi les liens avec les deux autres composants.

Searching for an intuitive way to understand natural illumination, chapter 3 dealt with the problem of sky classification based on its appearance. We explained that, in the case of

outdoor images, global illumination is highly determined by the atmosphere that surrounds the scene. The proposed approaches allowed us to group similar photos using different image analysis techniques. However, our algorithms use only a part of the information contained in the image. Important clues such as object appearance, geometry of the scene or shadows are discarded. Furthermore if we have access to a collection of images of the same scene, the number of variables shared by all the images might help to reduce the complexity of the problem. Thus, the estimation of the parameters defining the image formation process should become a doable task. In this chapter we explore how to reach this objective, by focusing our attention on the parameters defining the Camera Response Function (CRF), a non–linear curve that maps the scene irradiance into image intensities (*cf.* 2.1.4).

This chapter describes some approaches to estimate the CRF's along with the illumination and the object's reflectance properties in an increasingly complexity order. We start by exploring a simple case, when a planar surface is projected into two images observing the same scene. The two images shared a common planar surface. The formulation of this simple case is the basis for the oncoming approaches and it establishes the foundations of our methods. Next we introduce a more complex case and we propose a solution for multiple images sharing one planar surface. The natural extension of this approach brings us to generalize the solution for convex surfaces. Finally we propose a non–linear estimation using a simple model for the illumination: a directional light source. Results of the proposed methods on synthetic and real images are shown and a final discussion section concludes the chapter.

4.1 Linear Methods

4.1.1 The Two Images, One Plane Case

In section 2.1, an image formation model was introduced. Under some circumstances, the presented model can be simplified. For example, assuming that the object placed in the scene has Lambertian reflection properties, is a common approximation. Even if this assumption is not completely true (objects in the real world are neither completely Lambertian nor 100% specular), it is known that a large component of the perceived color is due to the Lambertian influence. Furthermore specular and semi–specular regions can be treated as particular cases and, their influence can be reduced using algorithms robust to particular and isolated situations. In the case of Lambertian objects, we showed that equation (2.5) describes the image irradiance $E_k(\mathbf{x})$ for a given surface point with spatial coordinates $\mathbf{x}$ and a camera sensitivity represented by q_k. Additionally, let us recall that under the assumption of a single directional light source ($\mathbf{L}_{ki}$), the image irradiance (E_{kji}) emitted by a point j on the surface of the object, also called the j^{th} *surface element*, can be denoted by (*cf.* equation 2.7):

$$E_{kji} = \frac{\rho_{jk}}{\pi} \cdot q_k \left(\mathbf{N}_j^{\mathsf{T}} \mathbf{L}_{ki} \right) \quad . \tag{4.1}$$

We assume here that there is only one illumination per camera at the same time, and that the process is equivalent for the three filters composing the camera sensor ($k \in \{R, G, B\}$). Then, for the moment, we can ignore the index k and redefine the camera sensitivity q_k as q_i:

$$E_{ji} = \frac{\rho_j}{\pi} \cdot q_i \left(\mathbf{N}_j^\mathsf{T} \mathbf{L}_i\right) \quad . \tag{4.2}$$

After this change, the index i refers to the illumination and the camera's properties as well, keeping in mind that for each camera and for each light vector $\mathbf{L}_i$ there are three different color channels (red, green, blue).

To get the intensity of the pixel representing the projection of the surface element, we include the camera response function into the image formation model (eq. (2.8)). Hence, the intensity shown by the surface element j on the image taken with camera i is expressed as:

$$B_{ji} = f_i \left(\frac{\rho_j}{\pi} \cdot q_i \left(\mathbf{N}_j^\mathsf{T} \mathbf{L}_i\right) \right) \quad , \tag{4.3}$$

and the albedo ρ_j for the j^{th} surface element corresponds to:

$$\rho_j = \frac{f_i^{-1}(B_{ij})\pi}{q_i \left(\mathbf{N}_j^\mathsf{T} \mathbf{L}_i\right)} \quad . \tag{4.4}$$

In section 2.1.4 we described a model for the Camera Response Function using an empirically-found basis of functions and some coefficients (*cf.* 2.9). There exists also an inverse model for this function, defined by $g(B) = f^{-1}(B)$. It has similar characteristics to the direct model previously described. The inverse model can also be represented as a linear combination of some basis functions and their corresponding coefficients:

$$g(B) = p_0(B) + \sum_{n=1}^{N} u_n p_n(B) \quad , \tag{4.5}$$

where the basis formed by $\{p_0(B), p_1(B), \cdots, p_N(B)\}$ corresponds to the inverse empirical model introduced by [Grossberg 2004].

Let us consider the scheme presented in figure 4.1. In this case we have two images captured using two different cameras. During the acquisition of image 1, the light source illuminates the scene from L1 and for the image in camera 2 the only active illumination is coming from L2. For a selected point on the surface of the object, its projections onto the image planes 1 and 2 reveal different color intensities. However the two images share the same surface reflectance properties. Hence,

$$\rho_j = \frac{g_1(B_{1j})\pi}{q_1 \left(\mathbf{N}_j^\mathsf{T} \mathbf{L}_1\right)} = \frac{g_2(B_{2j})\pi}{q_2 \left(\mathbf{N}_j^\mathsf{T} \mathbf{L}_2\right)} \quad . \tag{4.6}$$

Next, we will assume that the spectral response of the camera's sensor is the same for both images ($q_1 = q_2$).

$$\frac{g_1(B_{1j})}{\mathbf{N}_j^\mathsf{T} \mathbf{L}_1} = \frac{g_2(B_{2j})}{\mathbf{N}_j^\mathsf{T} \mathbf{L}_2} \quad . \tag{4.7}$$

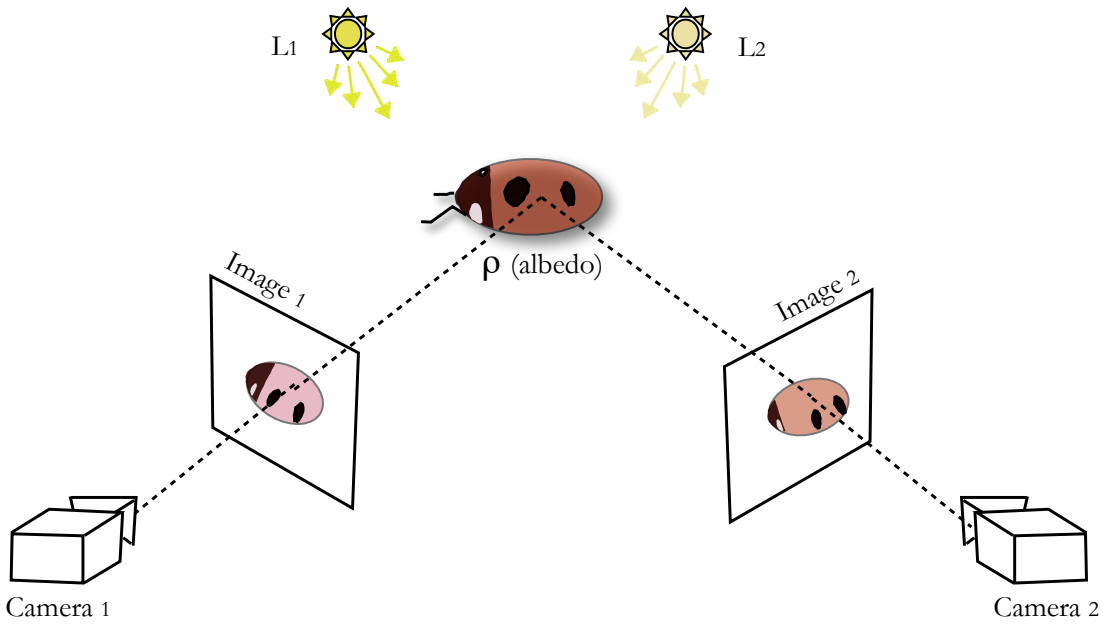


Figure 4.1: The two images case.

Also, to show the feasibility of our approach, the object in the 3D world is limited to be a plane. In this case, the whole surface is modeled by its normal. The dot product of the normal and the light is reduced to a cosine term plus the ambient term. We arrived at the following equality:

$$\frac{g_1(B_{1j})}{\cos\alpha_1 + \mu_1} = \frac{g_2(B_{2j})}{\cos\alpha_2 + \mu_2} \quad , \tag{4.8}$$

where the j^{th} point (surface element) belongs to the mentioned plane. Let us denote the cosines of the angles and the ambient term for the light as a_1 and a_2:

$$a_1 = 1/(\cos\alpha_1 + \mu_1) a_2 = 1/(\cos\alpha_2 + \mu_2) \quad .$$

Then, the equality (4.8) is transformed into the equation (4.9):

$$\left(p_0(B_{1j}) + \sum_{n=1}^{N} u_{1_n} p_n(B_{1j})\right) a_2 - \left(p_0(B_{2j}) + \sum_{n=1}^{N} u_{2_n} p_n(B_{2j})\right) a_1 = 0 \quad . \tag{4.9}$$

Here the unknowns are the coefficients of the inverse CRFs g_1 ($\mathbf{u}_1 = [u_{1_1}, \cdots, u_{1_N}]^{\mathsf{T}}$) and g_2 ($\mathbf{u}_2 = [u_{2_1}, \cdots, u_{2_N}]^{\mathsf{T}}$), a_1 and a_2.

We can express the equation (4.9) in a matrix–form. The vector $\mathbf{B}_1$ of dimension $P \times 1$ stores of all the points belonging to the plane and projected on image 1. In the same way, we define the vector $\mathbf{B}_2$ for the corresponding points on image 2. $\mathbf{p}_{01}$ is the outcoming vector when applying the function p_0 on all the elements of the vector $\mathbf{B}_1$ and $\mathbf{p}_{02}$ is the result of

the same process on $\mathbf{B}_2$. We define also the matrices $\mathsf{P}_1 = [p_1(\mathbf{B}_1), p_2(\mathbf{B}_1), \cdots, p_N(\mathbf{B}_1)]$ and $\mathsf{P}_2 = [p_1(\mathbf{B}_2), p_2(\mathbf{B}_2), \cdots, p_N(\mathbf{B}_2)]$. These matrices are of size $P \times N$. Thus we obtain a complete system of linear homogeneous equations:

$$\begin{gathered}(\mathbf{p}_{01} + \mathsf{P}_1\mathbf{u}_1)\, a_2 - (\mathbf{p}_{02} + \mathsf{P}_2\mathbf{u}_2)\, a_1 = \mathbf{0} \ , \\ \begin{bmatrix}\mathbf{p}_{01} & | & -\mathbf{p}_{02} & | & \mathsf{P}_1 & | & -\mathsf{P}_2\end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ \mathbf{u}_1 a_2 \\ \mathbf{u}_2 a_1 \end{bmatrix} = \mathbf{0} \ . \end{gathered} \tag{4.10}$$

The system in equation (4.10) can be solved by any least squares method. In this case, it is possible to find the estimations for a_1 and a_2 up to scale, but the coefficients for the inverse CRF (u_j) are found without ambiguity.

4.1.2 The Multiple Images, One Plane Case

For the multiple images case, one alternative consists in creating a sparse linear system computing all pairs of images with respect to one reference image. For example, if the selected camera generates image 1, the linear system to solve is expressed with respect to image 1 as follows:

$$\begin{bmatrix} \mathsf{P}_1 & \begin{bmatrix} p_{02} & \mathsf{P}_2 \end{bmatrix} & \mathbf{0} & \dots & \mathbf{0} \\ \mathsf{P}_1 & \mathbf{0} & \begin{bmatrix} \mathbf{p}_{03} & \mathsf{L}_3 \end{bmatrix} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathsf{P}_1 & \mathbf{0} & \mathbf{0} & \dots & \begin{bmatrix} \mathbf{p}_{0n} & \mathsf{P}_n \end{bmatrix} \end{bmatrix} \begin{bmatrix} -\mathbf{u}_1 \\ \frac{a_2}{a_1} \\ \frac{a_2}{a_1}\mathbf{u}_2 \\ \frac{a_3}{a_1} \\ \frac{a_3}{a_1}\mathbf{u}_3 \\ \vdots \\ \frac{a_n}{a_1} \\ \frac{a_n}{a_1}\mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{p}_{01} \\ \mathbf{p}_{01} \\ \vdots \\ \mathbf{p}_{01} \end{bmatrix} \ . \tag{4.11}$$

Solving this system (4.11), we can find the coefficients to model all CRF's and the ratios between the cosines of the angles formed by the plane normal and light sources (plus the ambient term). However, if one (or all) the images are perturbed with noise, this method can yield a bad approximation. In that case, we propose a final optimisation stage, using as initialization points the results of equation (4.11) (that allows us to speed up the convergence and to constrain the problem).

4.1.3 A General case: Multiple Images, Convex surfaces

Let us recall the equation (4.7), where we found how to express the albedo using information available in two images with different illuminations. If the surface of the object inserted in the scene is not a plane then each surface element has a different normal. The equality is expressed as:

$$g_1(B_{1j})\mathbf{N}_j^\mathsf{T}\mathbf{L}_2 = g_2(B_{2j})\mathbf{N}_j^\mathsf{T}\mathbf{L}_1 \ .$$

In full detail:

$$\left(p_0(B_{1j})\mathbf{N}_j^\mathsf{T}\right)\mathbf{L}_2 + \sum_{n=1}^{N}\left(p_n(B_{1j})\mathbf{N}_j^\mathsf{T}\right)(u_{1n}\mathbf{L}_2) = \left(p_0(B_{2j})\mathbf{N}_j^\mathsf{T}\right)\mathbf{L}_1 + \sum_{n=1}^{N}\left(p_n(B_{2j})\mathbf{N}_j^\mathsf{T}\right)(u_{2n}\mathbf{L}_1) \quad , \quad (4.12)$$

where we grouped known and unknown entities respectively. This equation is linear in the following unknowns: $\mathbf{L}_1, \mathbf{L}_2, u_{1n}\mathbf{L}_2, u_{2n}\mathbf{L}_1$, for $n = 1 \cdots N$. Let $\mathbf{q}_{nij} = p_n(B_{ij})\mathbf{N}_j$. Then, if we consider J surface elements, the associated equations (4.12) can be grouped together in the following matrix equation:

$$\begin{bmatrix} \mathbf{q}_{021}^\mathsf{T} & \mathbf{q}_{121}^\mathsf{T} & \cdots & \mathbf{q}_{N21}^\mathsf{T} & -\mathbf{q}_{011}^\mathsf{T} & -\mathbf{q}_{111}^\mathsf{T} & \cdots & -\mathbf{q}_{N11}^\mathsf{T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_{02J}^\mathsf{T} & \mathbf{q}_{12J}^\mathsf{T} & \cdots & \mathbf{q}_{N2J}^\mathsf{T} & -\mathbf{q}_{01J}^\mathsf{T} & -\mathbf{q}_{11J}^\mathsf{T} & \cdots & -\mathbf{q}_{N1J}^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{L}_1 \\ u_{21}\mathbf{L}_1 \\ \vdots \\ u_{2N}\mathbf{L}_1 \\ \mathbf{L}_2 \\ u_{11}\mathbf{L}_2 \\ \vdots \\ u_{1N}\mathbf{L}_2 \end{bmatrix} = \mathbf{0} \ . \quad (4.13)$$

This is a linear homogeneous equation system in $8(N+1)$ unknowns (as a reminder, N is the number of principal components in the prior on CRF's, in our work it is equal to 3). It can be solved to least squares using *e.g.* a singular value decomposition of the design matrix. The unknowns are, as explained above, estimated up to scale only. In order for the solution to be unique (up to scale), the normals of the available surface elements need to be sufficiently diverse; for example, if all normals are identical, *i.e.* if the scene is planar, then the above equation system is underconstrained. In that case, one can fall back to the plane-specific approach of the previous section (see 4.1.2). Also, the observed brightness values need to be sufficiently diverse. It may be possible to establish necessary conditions for the well-posedness of the problem.

Once the equation system is solved, the individual unknowns are estimated as follows. The lighting coefficients $\mathbf{L}_i$ are directly given (up to scale) by the solution of the system. The

inverse CRF coefficients u_{in} can be easily computed: let $\mathbf{U}_{in}$ be the estimated solution of $u_{in}\mathbf{L}_{3-i}$ $(i = 1, 2)$. Then, the least squares solution of u_{in} is:

$$u_{in} = \frac{\mathbf{L}_{3-i}^{\mathsf{T}}\mathbf{U}_{in}}{\mathbf{L}_{3-i}^{\mathsf{T}}\mathbf{L}_{3-i}} \quad .$$

Note that although the $\mathbf{L}_i$ are estimated up to scale only, there is no scale ambiguity on the CRF coefficients u_{in}. Finally, the albedos ρ_j can be computed using equation (4.6), *e.g.* by averaging the estimates coming from the two images.

If M images are considered, a linear solution of all unknowns is possible by simply stacking equations (4.13) for all pairs of images on top of each other, giving a large but highly sparse equation system. In the most general case, where all pairs of images have visible scene elements in common, there are a total of $4M(1 + N(M - 1))$ unknowns $\mathbf{L}_i$ and $u_{in}\mathbf{L}_i$. If all J surface elements are visible in all images, then there are $JM(M - 1)/2$ equations. A number of $J \geq 8(N + 1)$ surface elements is in general sufficient for a unique solution up to scale (a tighter bound on J is possible but complicated).

This approach becomes intractable if M becomes of the order of several hundreds or larger. In that case, we only use equations linking one reference image with all the other images.

4.2 Non Linear Methods: A Directional Light Source

The above linear methods are efficient (in the case of many images only if not all equations are used, however). In the presence of noise they are suboptimal, since the cost function that is minimized (the norm of the expression in equation (4.13)) is algebraic and the estimated unknowns constitute a redundant parameterization of the actual unknowns of the problem. An optimal solution thus requires the full non-linear optimization of a meaningful cost function (the function gives the maximum likelihood estimate if observed brightness values are affected by i.i.d. Gaussian noise). In this case, we aim at estimating the unknowns by minimizing the difference between observed and predicted brightness values. Let us suppose M is the total number of images and J is the number of surface elements of a 3D model. We seek to minimize:

$$\underset{\mathbf{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{M}\sum_{j=1}^{J}\left(B_{ij} - v_{ij}\hat{B}_{ij}(\mathbf{\Theta})\right)^2 \quad , \tag{4.14}$$

where the scalars v_{ij} are booleans, a value of 1 indicating that surface element j is visible in image i, otherwise the value being 0. In this equation, the predicted brightness ($\hat{B}_{ij}$) is calculated by using the image formation model with a single directional light source, as illustrated in equation (4.3). We shall incorporate the term q_i to the light vector $\mathbf{L}_i$ in such a case $\mathbf{L}'_i = \mathbf{L}_i q_i$ and the albedo $\rho'_j = \rho_j/\pi$. Thus, our estimation function is:

$$\hat{B}_{ij} = f_i\left(\rho'_j\left(\mathbf{N}_j^{\mathsf{T}}\mathbf{L}'_i\right)\right) \quad . \tag{4.15}$$

The function f_i is given by the direct model proposed by Grossberg and Nayar [Grossberg 2003], previously discussed and shown in equation (2.9). Hence, the estimation function is transformed into:

$$\hat{B}_{ij} = h_0\left(\left(\rho'_j\left(\mathbf{N}_j^{\mathsf{T}}\mathbf{L}'_i\right)\right)\right) + \sum_{n=1}^{N} w_n h_n\left(\left(\rho'_j\left(\mathbf{N}_j^{\mathsf{T}}\mathbf{L}'_i\right)\right)\right) \quad , \tag{4.16}$$

Note that the unknowns ($\mathbf{\Theta} = \{\rho', \mathbf{L}', [w_1, \cdots, w_N]\}$ can not be estimated without ambiguity: albedos ρ'_j and lighting coefficients $\mathbf{L}'_i$ can only be estimated up to one global scale factor. In contrast, the solution for CRF coefficients is unique, like in the previous section. In addition to the cost function (4.14), we also impose a monotonicity constraint on the estimated CRF's.

The Jacobian matrix and gradient of the least squares problem (4.14) are trivial to compute (an analytical solution for a similar case is explored in detail in section 5.2.4) and the Jacobian matrix is highly sparse. Since plausible CRF's are monotonic, we also impose inequality constraints on the CRF coefficients, as follows: for 1024 equidistant irradiance values between 0 and 1, we enforce that the CRF is larger for each sampled irradiance value than for the next smaller one. Depending on the use or not of these constraints, we use either the `fmincon` or the `lsqnonlin` optimization functions of MATLAB, which make full use of the sparsity of the Jacobian matrix. Both of these functions use refined versions of the Levenberg-Marquardt algorithm described in section 2.3.

In our experiments, we used different methods to initialize the unknowns. Besides using the linear method of the previous section (4.1.3), we also initialized the CRF's as linear functions $f_i(E) = E$ or as the average CRF of Grossberg and Nayar's analysis: $f_i = h_0$.

If the linear method of the previous section is used for initialization, we first have to estimate the coefficients of the direct CRF's, from those of the estimated inverse CRF's. To do so, we simply generate sample pairs of brightness and irradiance values from the computed inverse CRF's g_i, for a dense equidistant sampling of the range of brightness values. For each image, we then fit a direct CRF to these samples, by estimating the coefficients w_{in} of (2.9).

4.3 Experiments and Results

To evaluate the accuracy and the performance of the approaches proposed in this chapter, we ran experiments on synthesized data and also in real images. For the latter case, in addition to the databases described in section 3.2, we created a set of images simulating the conditions of images found on Internet collections.

In the next paragraphs we shall explain how we verified our approach for the two and multiple images case using a synthetically–created plane. We also explain how to validate the general approach using real images. A discussion around how to obtain a *ground truth* and a benchmarking test are presented at theend of the chapter.

4.3.1 Synthesized Data

We created different classes of synthesized images. First, to test the approaches described in 4.1.1 and 4.1.2, we simulated images for a plane with spatially varying albedo and a single directional light source (see figure 4.2). The brightness values are the result of mapping the irradiance with randomly selected real–world CRF (like those shown in 2.4). We generate im-

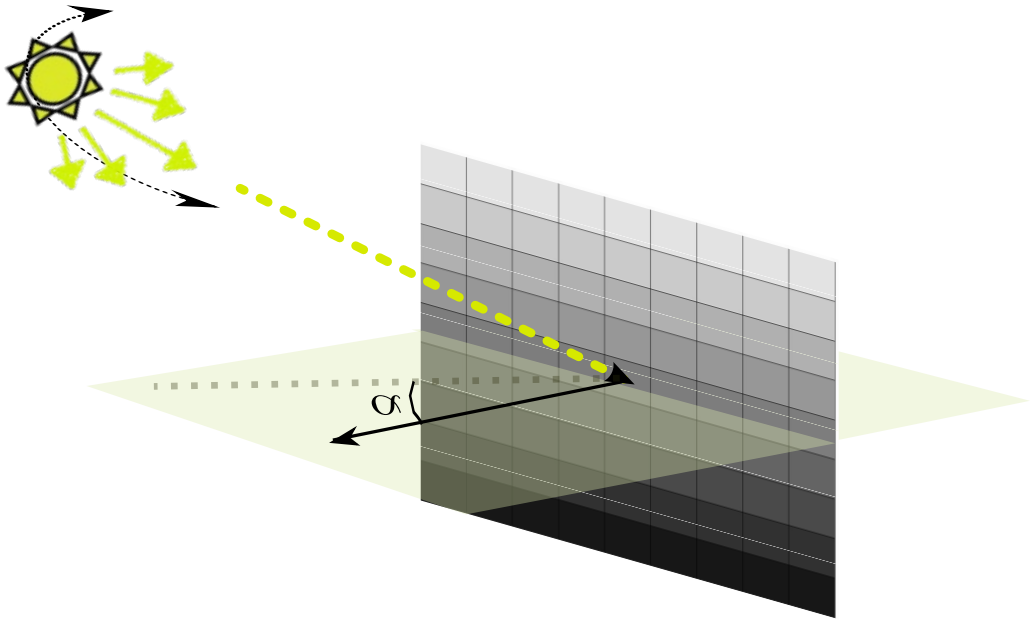


Figure 4.2: Synthesized plane images.

ages for different positions of the light source (*cf.* 4.8), varying the angle formed by the light and the plane normal (α) over a large range of possible values: $\alpha = \{10\,^\circ, 20\,^\circ, 30\,^\circ, \cdots, 80\,^\circ\}$. The ambient term μ in these experiments was equaled to zero. Examples of the images generated from this process for various configurations of angles and real–world CRF's are show in figure 4.3.

This set up allows us to check out in an experimental environment our analysis for the case of two and/or more images of a planar surface, taking into account the different models. We started by selecting images by pairs and by solving equations for the case of two images, one plane (4.10). Figure 4.4 shows the real and estimated inverse CRF's for a couple of randomly–selected synthesized images.

The angle α in one of the images, *i.e.* in image 1, was intentionally fixed to zero. Therefore, a large range of possible brightness values are available in this image. The system for different values of the angle formed by the plane and the light in second image (α_2). As seen in the figure, the estimated inverse CRF's are very accurate when the α_2 values are lower. On the other hand, when the angle α_2 increases the estimation fails because the full range of normalized brightness is not available. In this case the CRF can not be completely

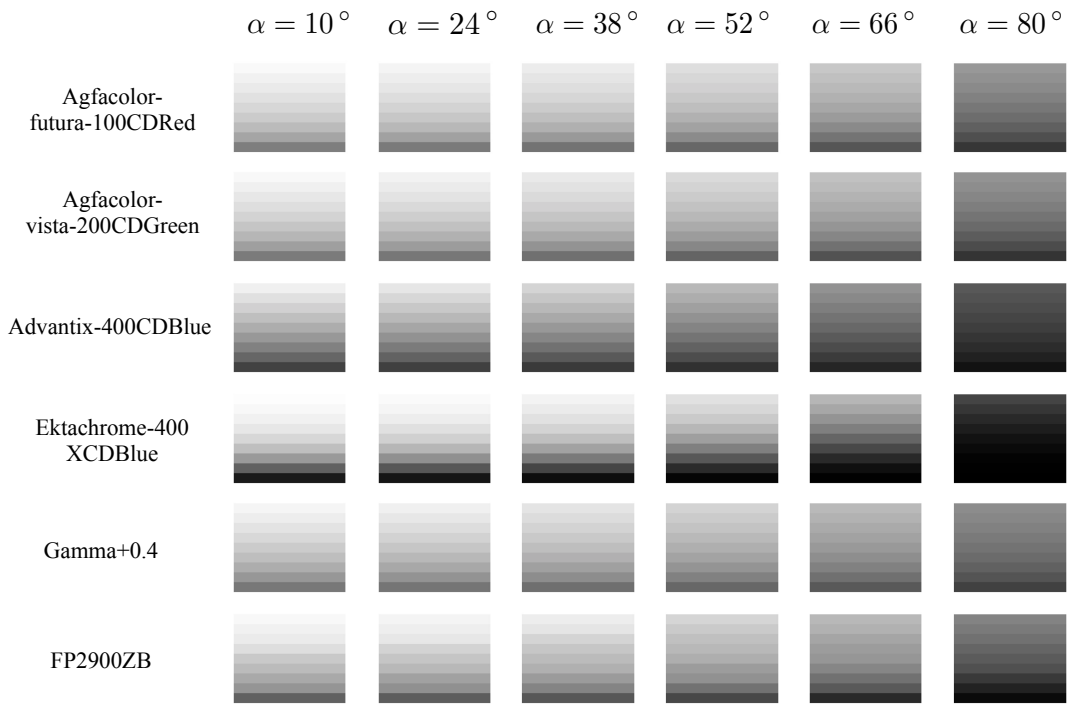


Figure 4.3: Synthesized plane images for different combinations of CRF's and angles α.

recovered. It is important to mention that angles greater than 40 degrees are not usual in real world environments. Moreover, the two images case is the simplest scenario and using multiple images should overcome some of these limitations. For the multiple image case, we solve equation (4.11) using 50 synthesized images. They were generated at random but relatively low light-plane angles in order to guarantee a large range of brightness available. Images were obtained using randomly selected CRF's, as described before. In this case we also add low levels of Gaussian noise to the synthesised images (the mean of the Gaussian noise is up to 0.5% of the maximum brightness value). We performed a final non-linear optimization step on the CRF coefficients. Figure 4.5 shows the original and the estimated CRF for six of the images resulting from this process. The non-linear optimization was initialized with the coefficients obtained from the linear solution (since the output of the linear approach corresponds to the inverse CRF, we found the direct coefficients by fitting them to the inverse function, as described earlier). Inclusion of noise in synthesized images affects drastically the performance of the linear approach. However the non–linear optimization improves the estimation. The estimated CRF's are very close to the original ones. We also show the *root mean squared error* (RMS) between the estimated CRF's and the original function. The RMS is calculated by taking the squared difference between the real measured CRF and its estimation over 1024 points (see equation (4.17)). In general, when dealing with noisy images, optimization improves the results of our algorithm.

A second kind of synthesized images were also created. In this case, they correspond

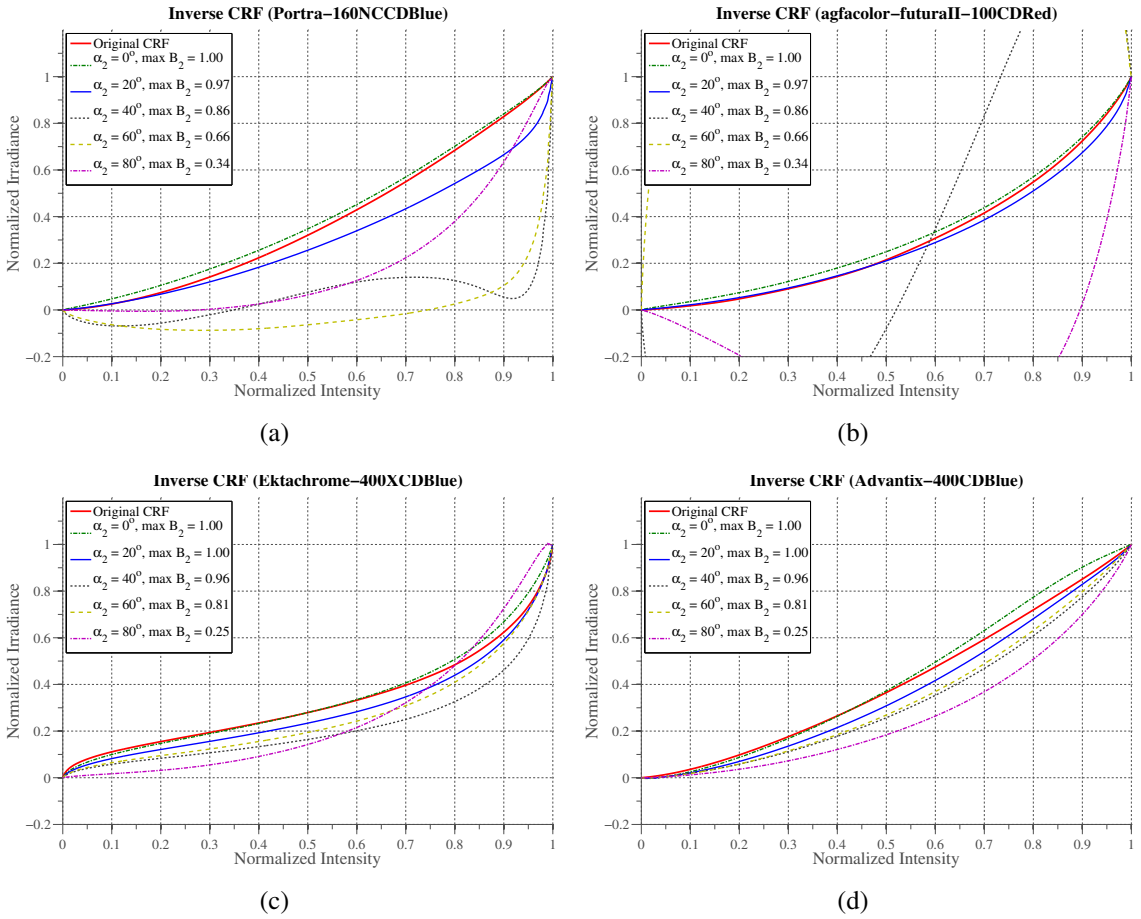


Figure 4.4: Plots (a) and (b) are estimated and original CRF's for camera 1 (g_1) and camera 2 (g_2) respectively, using different values of α_2 (α_1 is set to zero). Evident out-of-range values arise when α_2 increases and observed brightness values are not sufficient to model the curve. Plots (c) and (d) show the same estimation for a different pair of images synthesized with some of the real–world CRF's used in 4.3. All estimated curves use 3 coefficients to model the inverse CRF.

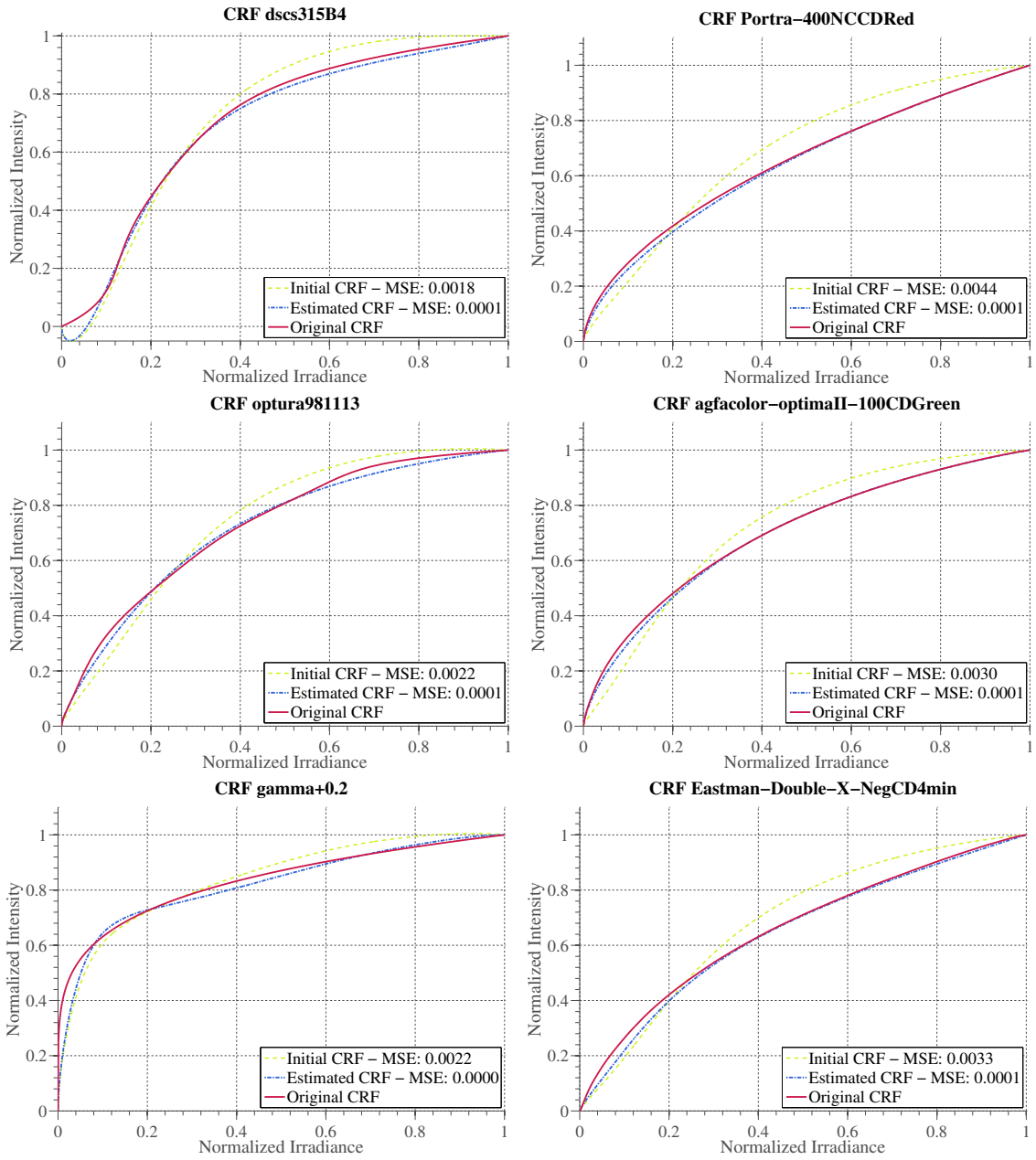


Figure 4.5: Estimated CRF's for six images (from a set of 50 images) before and after optimization.

to renderings of a sphere with spatially varying albedo. Each surface element belonging to the sphere is composed by the position of the 3D point plus its normal (see figure 4.6). 50 images of the sphere were produced. For each image, surface elements were exposed to different illuminations, *i.e.* different directional light sources and ambient lighting of different "strengths". To produce the respective brightness values, a randomly selected CRF was used for every image, as in the case of synthesized plane images (figure 4.7). These images are used to validate our approach when using multiple images and convex surfaces. The variety of the normals allows us to test the solution proposed in equation (4.13) and the non–linear optimization algorithm.

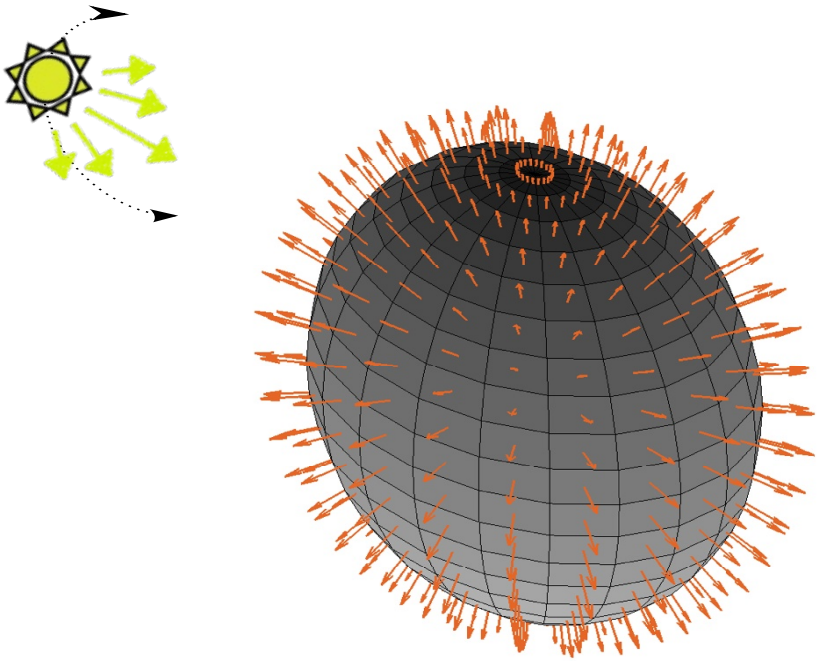


Figure 4.6: Sphere with spatially varying albedo. The colors present on the surface of the sphere correspond to the albedo without any light source contribution.

We added Gaussian noise to the generated brightness values, with standard deviations between 0.001 and 1 percent and a mean value of 1 percent of the range of brightness values. We ran 50 trials for each noise level. Figure 4.8 shows estimated and original inverse CRF's for 6 of the 50 cameras in a typical trial. The linear approach performs relatively poorly. The non-linear optimization however, even when initialized from the solution of the linear approach, leads to good results. A quantitative assessment is presented in table 4.1, for a varying noise level. The quality of an estimated CRF is measured by computing the following root mean square error (RMS) between the true and estimated inverse CRF:

$$\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(f(E_k)-\hat{f}(E_k)\right)^2}\ , \tag{4.17}$$

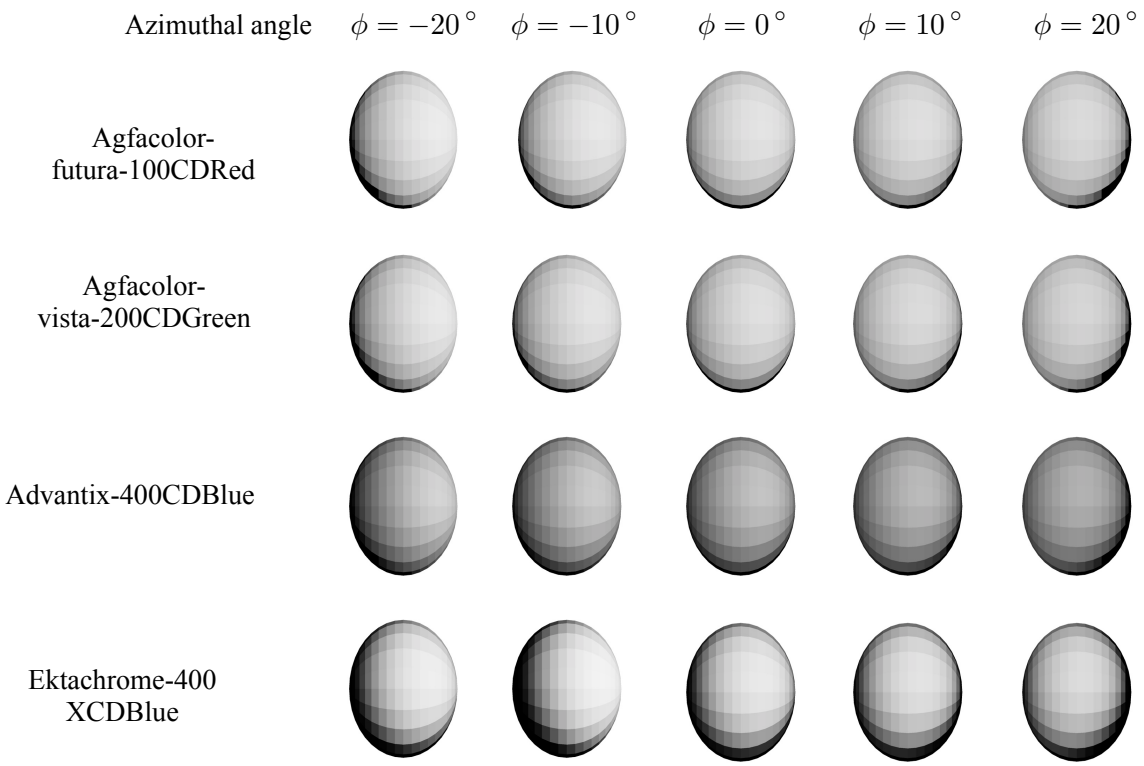


Figure 4.7: Synthesized sphere images for different combinations of CRF's.

Noise (St. Dev. [%])	0.001	0.01	0.1	1
RMS Error inverse CRF	0.032	0.040	0.044	0.043
RMS Error for albedo	0.073	0.098	0.112	0.138

Table 4.1: Median RMS error for 50 synthesized images

where the E_k are taken as the $K = 1024$ equidistant values between 0 and 1.

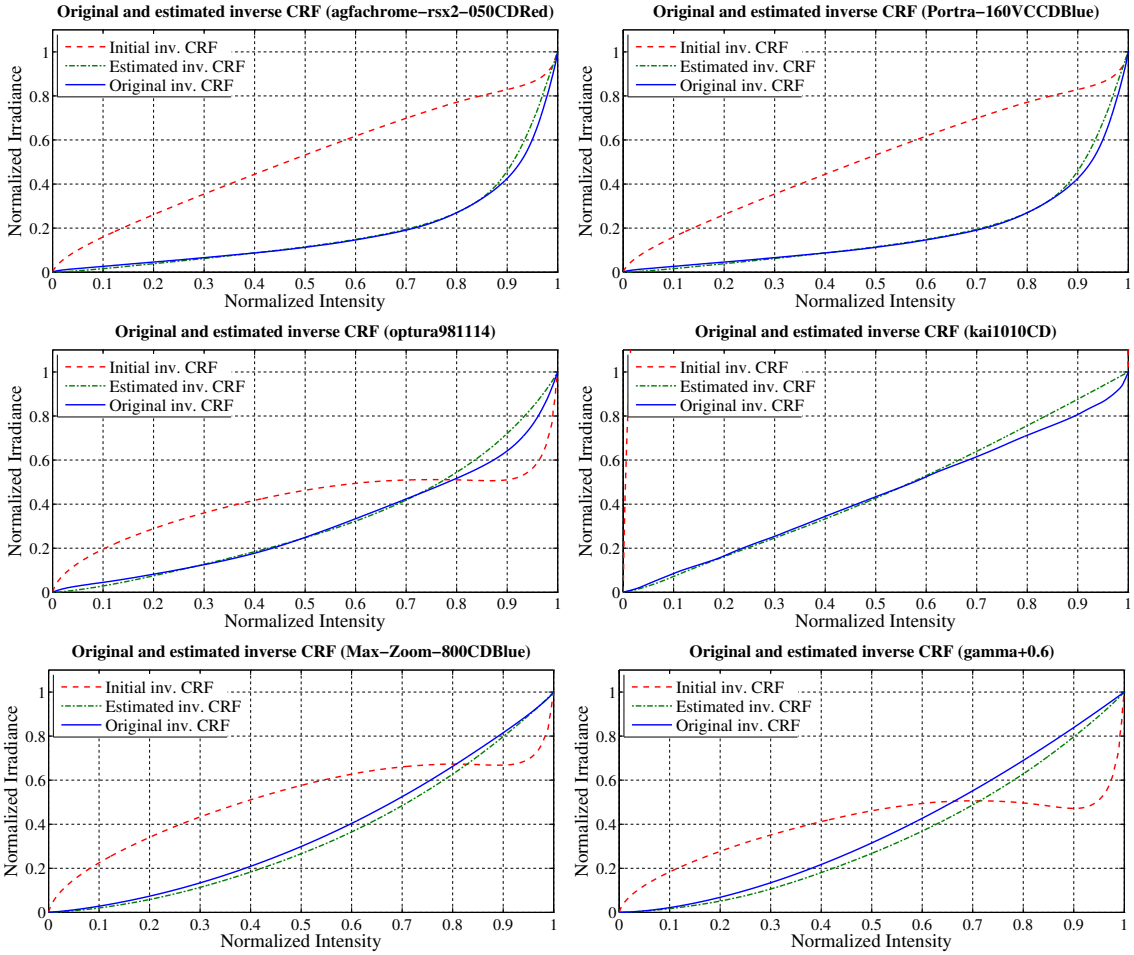


Figure 4.8: Six estimated inverse CRF's using 50 synthesized images perturbed by Gaussian noise of standard deviation 0.01 and mean 0.1 (*cf.* text). The initialization for the non-linear algorithm is given by the method of section 4.2. Non-linear estimation was taking into account monotonicity constraints on the CRF's.

4.3.2 Real Images

Manufacturers rarely provide technical information about the behavior of the captured intensities and how they are related to the irradiances of the scene. Moreover, following our definition of CRF, it implicitly includes changes induced by the environment, the optics, the camera set-up (white balance, exposure value, etc.), and internal processing of the data. As a consequence, a ground truth to test our algorithms is not readily available. Approximations to the ground truth can be obtained using samples of a color calibration chart, however controversy and discussion around the right RGB values for the chart's patches have been raised (see [Pascale 2006]). Additionally, as explained in section 2.1.4, without a knowledge of the photometric characteristics of the sensor, to estimate a reliable CRF is not an evident task. In the same section we mentioned some approaches to get good approximations for the CRF: most of them need physical acces to the scene. For that reason, we recreate our own database simulating the characteristics of an Internet photo collection. We took 120 photos of an outdoor scene with a set of 5 different cameras. Images were acquired during different periods of the year and under changing illumination conditions. Some samples of the images can be seen in figure 4.9. These photos were taken using an automatic setup for the cameras (white balance, aperture, exposition time, etc.). At the same time, for some particular positions, we also took multiple images with different exposures, without altering the camera pose. In those cases, just some seconds after the first shots, the color checker chart was also included in the scene, keeping, when possible, the camera settings locked between the original shot and the image containing the color chart. We are aware that the inclusion of the color chart could change camera parameters when shooting in automatic mode (*e.g.* white balance) then, when possible we change the mode to manual and we use the settings determined at the initial shoot. When possible, image where acquired in raw format and converted to JPG using the white balance information available in the metadata. This process was performed using the open source software *dcraw*[1]. Table 4.3.2 details the characteristics of the database: the camera models used for acquisition, the number of photos taken, the format for image files, etc.

Camera Model	No. Photos	File Form.	Mult. Exposure?	Color Chart?
Canon EOS 450D	45	RAW	×	×
Canon PowerShot DS10	25	RAW	×	×
Canon PowerShot A510	25	JPG	—	×
Sony DSC-H2	15	JPG	×	×
Iphone 3GS	10	JPG	—	×

Table 4.2: Details on the acquisition conditions for the database.

We also test the performance of our algorithm with images downloaded from the internet, using the photographs compiled for the database described in section 3.2. This time we use

[1] http://www.cybercom.net/~dcoffin/dcraw/

Figure 4.9: Sample images of the database simulating an Internet photo collection

958 images of the *Sacré Cœur's* Cathedral. In both cases, recovering the 3D geometric model is a preliminary necessary step. We explain this stage of our algorithm in the following paragraphs.

3D Geometric Model Reconstruction. Multi–view reconstruction of large scale scenes has progressed in a surprising way in recent years. We have already discussed some of the algorithms used in these type of problems. In this work, we recover the camera pose using the software called *Bundler*[2] described in [Snavely 2008]. The output of this system corresponds to the pose of the cameras given by their rotation matrix and position, along with the values for the intrinsic parameters. Initial guesses for the focal lengths and the pixel densities are extracted from the images metadata. Along with the camera's poses, a sparse reconstruction is also provided. It is composed by the position in the 3D space of the feature points used for the pose estimation. This algorithm also computed a radial distortion parameter and corrects the images suffering from this kind of distortion. The named outputs are the starting point for a multi-view stereo reconstruction algorithm. We used the code of *CMVS*[3] provided by Furukawa *et al.* [Furukawa 2010]. Using a novel partition of the 3D space, this software creates a dense reconstruction of the 3D scene, represented by a point cloud. Each point is represented by a position and a normal. While for some experiments we use directly the output of this algorithm (the 3D points and their normals), in others we create a mesh from the point cloud using the Poisson reconstruction algorithm described in [Kazhdan 2006]. In those cases, a final manual refinement step was performed on the mesh. Recovered 3D models for the case of point clouds and meshes are shown in figure 4.10

Using the ColorChecker® chart. The standard approach to estimate the CRF usually places an object of known radiance in the scene during the acquisition. The company *X-Rite Photo* manufactures a chart commercially called the ColorChecker® Classic. It consists of a target containing 24 patches which represent the actual color of natural objects and reflects light like in real-world scenarios. The colorimetric data for each of the 24 patches is given by the chart's manufacturer in standard color appearance models ($sRGB$ and CIE La^*b^*). To assure the success of the calibration, the user must guarantee that the image color pixels are in the same color space as the colorimetric data. As we mentioned before, we include the color chart during the acquisition of some images of the previously described database. Some samples of these images are shown in figure 4.11. When raw images were obtained, they were converted to the $sRGB$ color space using the white balance information given by the EXIF metadata. In the other cases we verified in the camera specifications that the pixel values are given in the right color space. We manually selected rectangular regions inside each patch and the mean value of the pixels in the selected area was calculated. Once we get the 24 pairs of irradiance–intensity values, the CRF was computed by fitting the empirical

[2] http://phototour.cs.washington.edu/bundler/
[3] http://grail.cs.washington.edu/software/cmvs/

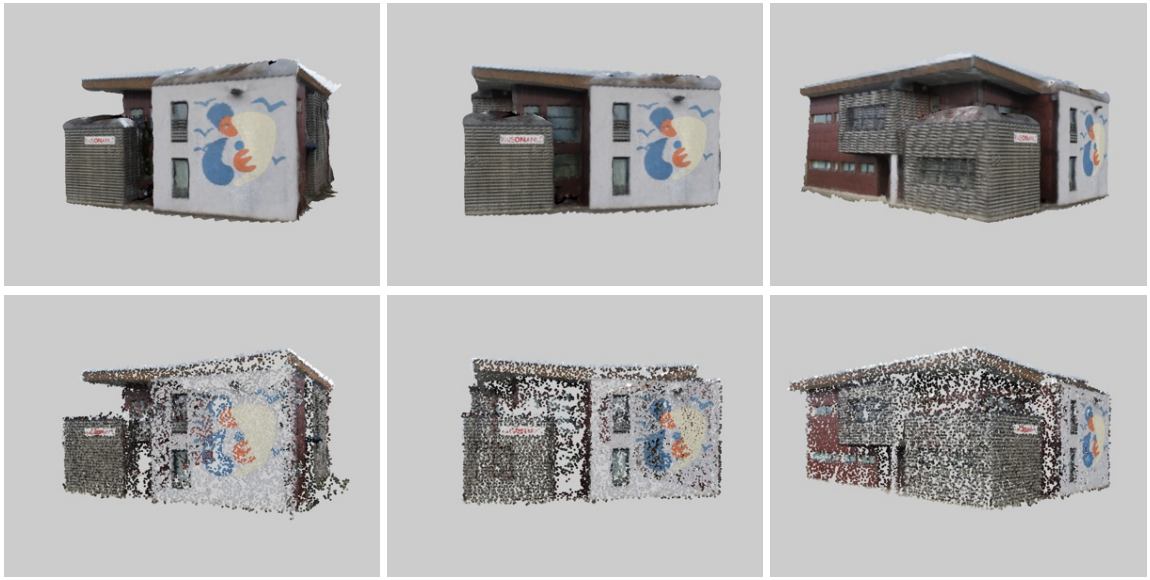


Figure 4.10: Example of the recovered 3D models: mesh and point cloud representations. Colors in both representations correspond to the mean value of pixels corresponding to the surface elements on the images where they are visible.

model of Grossberg and Nayar [Grossberg 2004] to the samples. These results are then considered as the *ground truth* for the benchmarking test explained later.

Figure 4.11: Example of images showing the ColorChecker in the scene during the acquisition.

Multiple exposure methods. In section 2.1.4 we mentioned the multiple exposure methods among the active approaches commonly used to estimate the CRFs. Indeed, Devebec and Malik [Debevec 1997] were pioneers of high dynamic range (HDR) imaging when they employed the estimation of the response function to increase the range of colors captured by a camera. In their method, authors propose to use differently exposed photographs to recover the response function of the imaging process. The implementation of this algorithm is

available through the *HDRShop*[4] software available for "non–commercial purposes" on the web. In [Mitsunaga 1999], Mitsunaga and Nayar also use the multiple exposures approach to find the CRF. In that case, authors assume that the camera function can be modeled using a high order polynomial: $B = f(E) = \sum_{n=0}^{N} c_n E^n$. Grossberg and Nayar also propose to compute the CRF from multiple exposure images [Grossberg 2003], but using their proposed empirical model for the response functions. The common point of all the mentioned methods is that they suppose that image irradiance is proportional to the ratio of exposure between perfectly aligned images. For example, if two images of the same scene with intensity values B_1 and B_2 are captured with varying exposures e_1 and $e_2 = k \times e_1$, where k is the ratio of exposures, then irradiances estimated from the inverse CRF must satisfy $g(B_1) = kg(B_2)$. To proceed with our comparison test, we implemented a version of the latter mentioned algorithm [Grossberg 2003] by solving the linear equation system generated when using 6 images taken at different exposures (f-stop number varying in steps of $1/3$). Figure 4.12 illustrates a set of multiple exposure images captured during the acquisition process.

Figure 4.12: Multiple exposure aligned images used to estimate the CRF.

Passive CRF estimation: a single image. The methods for estimating the CRF described before have a strict constraint, hard to avoid. The user must have the camera in hand, or at least, readily available. Lin *et al.* propose an algorithm to work around this limitation by exploiting automatically selected edge information to calculate the CRF. Using the empirical model of Grossberg and Nayar [Grossberg 2003], their method employs a Bayesian approach to formulate a minimization function in order to calculate the coefficients that describe the camera function. The algorithm works like this: Edge regions presenting two

[4] http://gl.ict.usc.edu/HDRShop/

uniform color distributions are selected, using image analysis techniques (edge detection, morphology, statistics for image regions, etc). Once a set of possible regions has been validated, the color pixels for the two regions and some intermediate points are ranked. They form the *observation set*. In ideal conditions, the path joining the colors from region 1 to region 2, passing through the intermediate points in the RGB color space must describe a straight line. However, given the non–linearity of the sensor, this path reflects the curvature induced by the CRF. Using a Bayesian estimation framework, authors propose to minimize the following function, in order to estimate the coefficients $\mathbf{u}$ of the optimal inverse CRF g^* (*cf.* equation (4.5)):

$$g^* = \min_{\mathbf{u}} \lambda D(g(\mathbf{u}); \Omega) - \log p(g(\mathbf{u})) \ , \tag{4.18}$$

where D is a function measuring the "straightness" of the path joining the pixels stored on the observation dataset Ω and $p(g(\mathbf{u}))$ describes the probability of having an inverse CRF with coefficients $\mathbf{u}$. This prior probability is found by using the empirical model of Grossberg and Nayar. Finally the parameter λ is not well described by the authors, but it seems to be a regularization parameter related to the number of regions selected for the estimation; it is set empirically to 10^4.

Benchmarking. We compare our results with some of the methods previously described. Using the multiple exposures images, we are able to compute the estimated CRF using the *HDRShop* software [Debevec 1997] and the multiple exposures method described by Grossberg [Grossberg 2003]. CRF estimation using our own implementation (with help by the authors) of Lin's *et al.* algorithm [Lin 2004] is also calculated.

We ran our algorithm on a set of 21,021 surface elements given by the 3D reconstruction process; here we use 3D points plus normals. Results for 6 image-camera pairs are shown in figure 4.13. Our estimation lies most of the time very close to the curve fitted to the color checker samples. Lin's method also exhibits good results, however best results are not always obtained with the recommended value of 10^4 for the involved regularization parameter λ [Lin 2004], see *e.g.* the lower right graph of figure 4.13. The upper right graph shows an example where Lin *et al.* 's approach gives a result relatively far from the ground truth.

Multiple exposure methods present irregular curves in some estimations. This can be explained because it is almost impossible to get perfectly aligned images when shooting outdoor scenes, even if the time between each shot is shorter than one second. Shadows, reflections and other artifacts can change rapidly, and success of these methods depends on perfect pixel alignment.

We also tested the algorithm using a database of images harvested from the Internet. Once we reconstructed the 3D model of the scene, 21,612 surface elements were available. To estimate the CRF, we extracted a random selection of 103 cameras and ran our algorithm (we do not use all images in order to reduce processing time and complexity). Because in this case it is impossible to collect samples using a color calibration chart neither to obtain multiple exposures with a static camera, we can only compare our results with the estimations

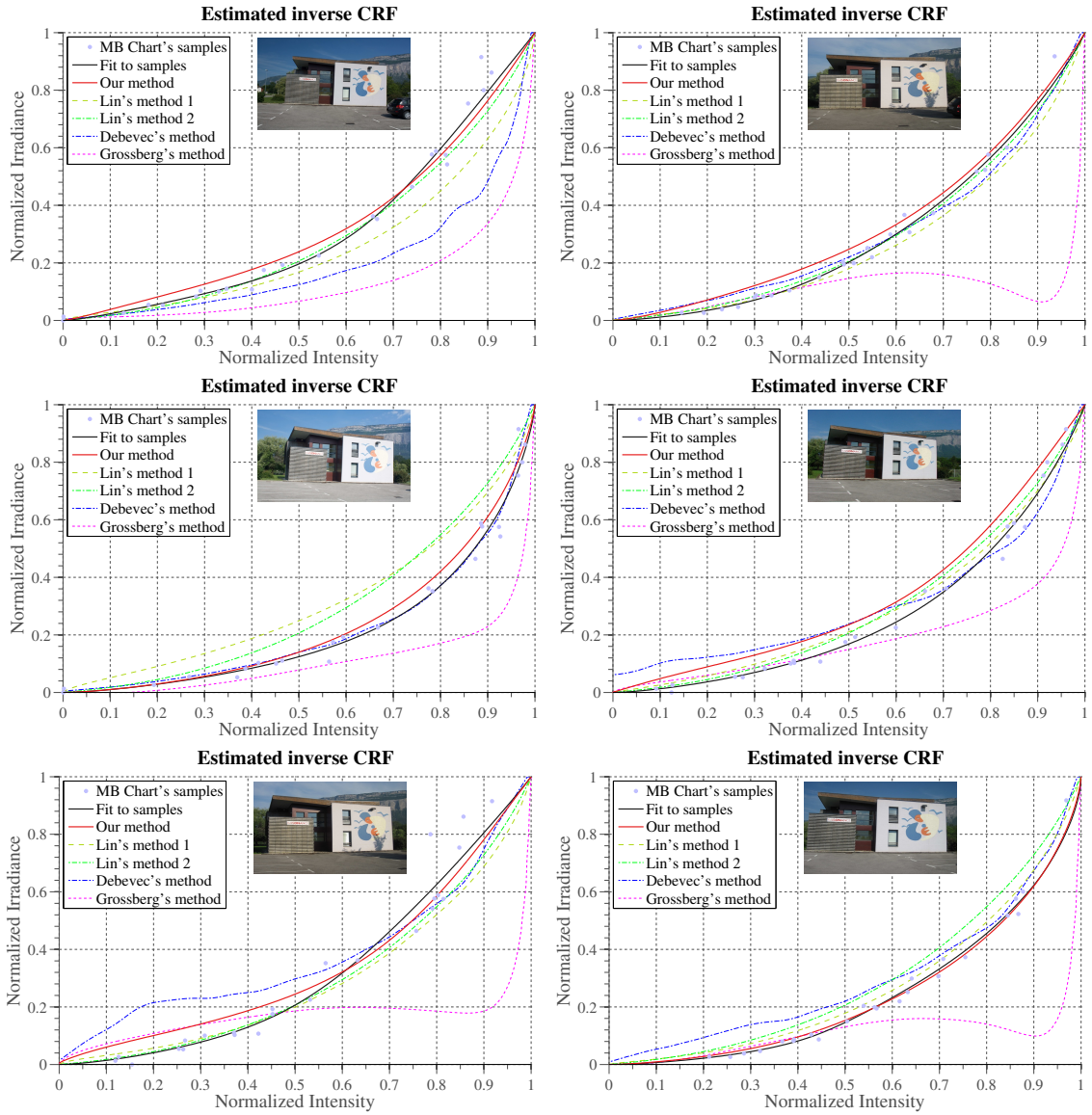


Figure 4.13: Estimated inverse CRF's found with the algorithm described in section 4.2 and compared to: Measured inverse CRF's with the color checker samples interpolated. Lin's *et al.* algorithm using two different values for parameter $\lambda = \{10^1, 10^4\}$ (*cf.* [Lin 2004]). HDRShop results [Debevec 1997]. Grossberg's *et al.* multiple exposures approach [Grossberg 2003]. Constraints in the monotonicity of the estimated CRF's are imposed.

RMS error/Database	Channel R	Channel G	Channel B
1. Own database	0.0431	0.0528	0.0463
2. *Sacré Cœur* Database	0.0627	0.0716	0.0679

Table 4.3: Median of the RMS error for the two databases described in section 4.3.2. RMS error is calculated relative to Lin's estimation (with fixed parameter λ) in both cases.

computed by Lin's algorithm. In figure 4.14, six recovered inverse CRF's are shown. Despite the dependence of Lin's method to the aforementioned regularization term, this approach has been to date one of the most reliable methods for estimating the CRF of single images. Our proposed solution is very close in most of the cases to the results of Lin's algorithm. We are also aware that Lin's method could present problems when it handles pre-processed images (the non-linearity of the CRF is altered along edges in this case). In table 4.3.2, we summarize the median of the RMS error over all the estimated CRFs relative to Lin's estimation.

4.4 Discussion and Conclusions

We have presented an approach to recover the camera response functions of a set of images in a photo collection, acquired with different cameras. The approach exploits a 3D model generated using available methods and a powerful empirical prior on CRF's, without which the problem would be ill-posed. Compared to other methods, our approach does not require multiple aligned images with the same CRF or the same lighting conditions. Our method is motivated as a first step in the pipeline of recovering surface properties and illumination conditions from multiple images and, in this context it can not be fairly compared with methods for single image CRF recovering [Lin 2004, Lin 2005, Matsushita 2007, Wilburn 2008, Kim 2008, Takamatsu 2008] (even so, results are as good as those estimated by these algorithms). Moreover, our method estimates all the CRFs for a group of images at the same time, in an automatic way, while, according to our experience, single image methods require most of the time some supervision or a manual tuning of the parameters, depending on the input image. For the case of image collections obtaining the CRF with these methods could be an extremely slow and tedious task. In this context, our results are promising.

Limitations of our approach are as follows. The illumination model, consisting of a single directional light source and ambient lighting, is certainly not entirely realistic. More general models can in principle be used, *e.g.* ones based on spherical harmonics [Basri 2003, Ramamoorthi 2001a, Ramamoorthi 2002]. Also, empirical priors on outdoor lighting conditions seem to be a good alternative. This "spherical harmonics" framework opens the door to the formulation of priors based on complex BRDFs as illustrated in [Romeiro 2010]. The next chapter of this thesis will introduce some of these techniques in the present context in order to evaluate a joint estimation framework for the models describing the image formation

process.

The empirical camera model proposed by Grossberg provides a good approximation to CRF's of cameras, where there is no evidence that different color channels' CRF's are correlated. However, the joint estimation of CRFs over all channels generates consistent CRFs. Color information could be used in the future, changing the assumption of a similar sensor spectral distribution in order to take into account the influence of the light wavelength on the computations. Another limitation is the assumption of Lambertian reflectance, although this is relatively easy to circumvent by using a robust weighting function in the non-linear optimization and having recourse to iteratively reweighted least squares optimization for example. A more detailed presentation of this will be exposed in next chapter.

We currently do not handle shadows associated with the directional light source. This could also be taken care of to some degree by a robust weighting function. One may also handle shadows explicitly since from the known scene geometry and the current estimate of the light source, shadows can be predicted and used to correctly predict brightness values. However, the resulting cost function would become discontinuous and harder to optimize. An alternative would be to directly detect shadows in the image with dedicated methods such as [Finlayson 2006] and use the result as input to our methods.

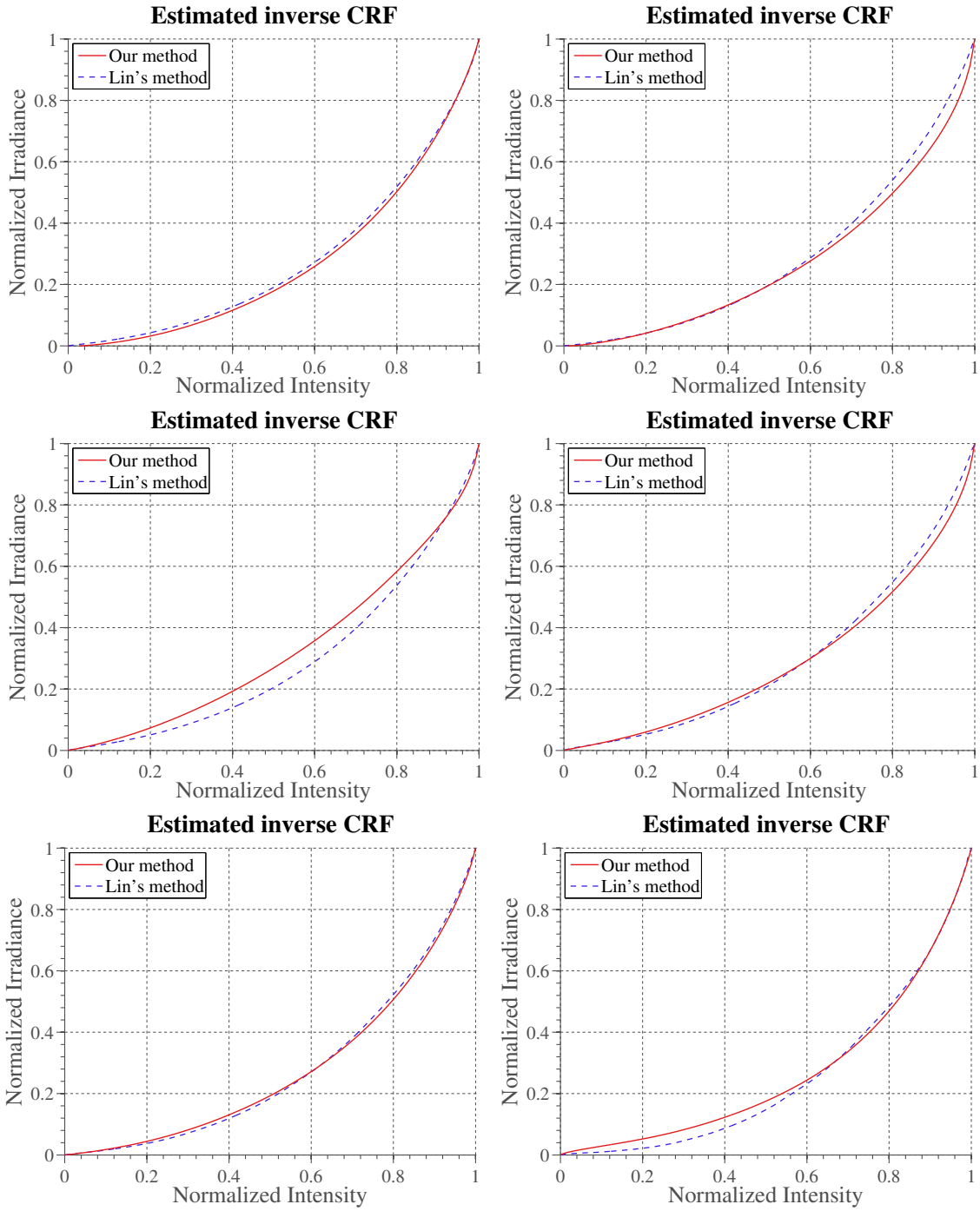


Figure 4.14: Estimated Inverse CRF's using our algorithm on a 3D model reconstructed from Internet images. Results of our own implementation of Lin's *et al.* [Lin 2004] work are also shown.

CHAPTER 5

Joint Estimation: CRF, Reflectance Properties and Global Illumination

Contents

Résumé. Les chapitres précédents nous ont permis de présenter des différentes méthodes pour extraire l'information photométrique à partir de collections d'images qui représentent des scènes d'extérieur. En particulier, le chapitre 4 a approfondi dans l'étude de la fonction non linéaire qui caractérise le capteur contenu dans des appareils photo numériques, formellement appelée la fonction de réponse de la caméra (CRF). La méthode proposée est basée largement dans le fait que les propriétés de la structure 3D restent invariables pendant l'acquisition des images. En utilisant des techniques de reconstruction multi–vues nous avons proposé de récupérer un modèle 3D de la scène, puis nous avons suggéré de modéliser les composants principaux qui prennent partie dans la formation de l'image. En effet, cet argument a été la base pour trouver le calibrage radiométrique des appareils photos qui ont pris les images de la collection. Cependant, dans la méthode présentée préalablement seulement quelques mots ont été mentionnés au sujet des résultats "résiduelles" du processus : Les paramètres qui décrivent un

modèle d'illumination simple et les propriétés de réflectance pour une surface hypothétiquement Lambertienne. Dans ce cinquième chapitre nous allons proposer une version améliorée du processus d'estimation, en utilisant des méthodes robustes pour le calcul des paramètres inconnus et un modèle plus détaillé pour l'illumination. Tout d'abord nous allons expliquer un modèle pour représenter une illumination global complexe en termes d'harmoniques sphériques. Dans ce cadre, l'interaction entre la lumière et la réflectance de la surface est carrément simplifiée. Ensuite, nous allons décrire comment résoudre le problème posé en utilisant une optimisation robuste non-linéal. Nous montrerons les similarités qui existent entre notre algorithme et la solution bien connue pour la reconstruction géométrique appelée Ajustement des Rayons (Bundle Adjustment). Nous incluons aussi une section en décrivant les caractéristiques de la matrice Jacobienne utilisée pour l'optimisation. Nous allons conclure le chapitre par une description des expériences menés.

In the previous chapters we have explored how to get photometric information about the scene and the cameras by exploiting outdoor image collections. Particularly, in chapter 4, we focused our efforts on the study of the non linear function that characterizes the camera sensors, formally called the Camera Response Function (CRF). The proposed estimation is based on the fact that the properties of the 3D structure visualized in the scene do not change during the acquisition of the images. These properties, *i.e.* the geometry and the surface reflectance properties, are common to all the images where the scene structure is projected. Given that nowadays it is possible to obtain a 3D geometric model from a set of outdoor images using multi–view stereo reconstruction techniques, the problem is reduced to retracing the image formation process and to estimate the parameters that govern the models describing this physical phenomenon. Indeed, in the last chapter we used this argument to successfully find the camera response functions. However, only a few words were mentioned about the "residual" results of the process: the parameters that describe the simple illumination model and the reflectance properties for an hypothetic Lambertian surface.

In the present chapter we propose an enhanced version of the process using a more detailed model for the illumination and a robust estimation of the parameters, focusing our attention not only on the CRF's but also on the lighting. We start by describing a framework to represent a complex global illumination in terms of spherical harmonics. In this context the interation of the light and the surface reflectance properties is clearly simplified. Next, we shall describe the robust non–linear solution for the proposed problem and we shall explore the similarities found between our formulation and the widely explored problem of Bundle Adjustment. We also include a brief section depicting the characteristics of the Jacobian matrix of the cost function computed during the non–linear optimisation. We conclude the chapter showing the results of our algorithm and proposing an open discussion on the advantages and disadvantages of our method.

5.1 Spherical Harmonics Illumination: A Global Illumination Framework

Finding object appearance under general illumination conditions seems to be an intractable problem. For example, a video projector could act as a light source, by illuminating an object with essentially any pattern. This extreme example shows us the great difficulties encountered when modeling lighting. Fortunately cases like that are not common when dealing with outdoor scenes in every–day scenarios, and in general the problem can be worked out by making a set of assumptions that can be easily reached. The first one consists in assuming a distant illumination source. This implies that the direction and intensity of the light source are the same along all the regions of interest. In outdoor scenes, this hypothesis is reasonable since the main light sources (sun and sky) are usually far away from the object.

In the case of exterior locations a second limitation concerning the sky illumination should be also treated. Indeed, we already mentioned it in section 2.2 when we presented models to represent the sky appearance. Most of the physical models stand up for illuminations coming from any incident direction (despite that the most probable directions correspond to the hemisphere situated above the object). Also the illumination can be composed of multiple light sources to be taken into account. In general, we would need to model the contribution of many lighting directions and the final result can be found by adding up their results. But in practice it is impossible to consider an infinite set of lights source to reach this goal. Fortunately there exists an approximation for the illumination acting in a scene under the mentioned circumstances. It is called environment mapping (*cf.* subsection 2.1.3). It can be seeing as a sampling of the infinite set of lights where it resolution (sampling rate) is choosen to get a good representation.

Based on the results of Ramamoorthi and Hanrahan [Ramamoorthi 2001a, Ramamoorthi 2001b, Ramamoorthi 2002], the interaction of the global illumination, with characteristics similar to the previously mentioned, and the BRDF can be expressed in simple terms within an appropriate framework. This result is supported by the analysis of the equation describing the scene radiance in terms of the spherical harmonics functions [MacRobert 1967, Groemer 1996]. The works of Cabral *et al.* [Cabral 1987] and D'Zamura [Zmura 1991] were pioneers on the use of spherical harmonics to analyze the behavior of the illumination and the BRDF. Basri and Jacobs [Basri 2003] show that the set of possible irradiances coming from a diffuse surface under a distant light source can be sufficiently approximated in a 9D linear space. This achievement was reached almost simultaneously than in Ramamoorthi's work, where similar results were presented. Recently more elaborate representations based on wavelets [Ng 2004] have allowed to analyze descriptions of general reflectance functions and natural lighting environments, however their complexity is a handicap to integrate it into our approach. In the next paragraphs, we shall detail the formulation of Ramamoorthi and Hanrahan and we shall include their results into our image formation model in order to propose a solution for our problem.

In the case of diffuse (or technically speaking, Lambertian) objects, the approach interprets the interaction of the illumination and the object reflectance model as a linear relationship, where the image irradiance can be seen as a filtered version of the incident illumination. The Lambertian BRDF acts as a low order filter. Experiments show that 9 spherical harmonics terms provide accurate results. To reach this result the main assumptions made by the authors of [Ramamoorthi 2001b] are:

- The object can be modeled by a convex Lambertian surface reflecting a distant illumination field. Since the object surface can be represented by its orientation, the surface normals describe the geometry of the structure.

- The reflectivity of each point on the surface can be characterized by an (potentially spatially varying) albedo.

5.1.1 The Image Formation Process viewed as a Convolution

In order to explain how to express irradiance as a linear combination of a basis and some coefficients, we shall review the contributions of Ramamoorthi and Hanrahan [Ramamoorthi 2001b] and how to adapt them to our problem. Let us recall the image formation equation for Lambertian surfaces described in equation (2.5). In that case, we mentioned that the image irradiance for the surface element localized in the coordinates $\mathbf{x}$ is related to the illumination L parametrized by the incoming angles θ_i and ϕ_i. These angles are measured with respect to the normal $\mathbf{n}$ at the point $\mathbf{x}$, thus we call these the local incident angles. In order to match our notation to the one used in Ramamoorthi's work, we are going to call the local incident angles θ_i' and ϕ_i' and we will reserve the unprimed variables to refer to the values linked to a global coordinate system:

$$E_{jk} = \frac{\rho_j}{\pi} \cdot q_k \int_{\Omega_i'} L_k(\theta_i', \phi_i') \cos\theta_i' d\Omega_i' \ . \tag{5.1}$$

For simplification purposes we will arrange the terms by denoting the albedo and the camera sensitivity in such a way that: $\rho_{jk}' = (\rho_j \cdot q_k)/\pi$. Under the assumption of a distant illumination, *i.e.* an illumination field homogeneous over the surface and independent from the surface position $\mathbf{x}$, the incoming light depends only on a global coordinate system and it can be parameterized by the *global* incident angles (θ_i, ϕ_i). Also, the surface can be reparameterized as a function of the normal of each surface element $\mathbf{n}_j$:

$$E_{jk}(\mathbf{n}_j) = \rho_{jk}' \int_{\Omega_i'} L_k(\theta_i, \phi_i) \cos\theta_i' d\Omega_i' \ . \tag{5.2}$$

The main idea behind Ramamoorthi's formulation is to find an expression analogous to the equation describing the spatial convolution operator. This objective is reached by expressing

the global coordinates in the equation (5.2) in terms of the local variables. Let us consider the normal vector parameterized by its spherical coordinates (α, β, γ). Then:

$$\mathbf{n} = [\sin\alpha\cos\beta, \sin\alpha\sin\beta, \cos\alpha] \ .$$

Consider also the transfer function $A(\theta'_i) = \cos\theta'_i$. With these modifications the last equation becomes:

$$E_k(\alpha_j, \beta_j, \gamma_j) = \rho'_{jk} \int_{\Omega'_i} L_k(\theta_i, \phi_i) A(\theta'_i) d\Omega'_i \ . \tag{5.3}$$

To find the relation between the global and the local coordinates authors define a rotation operation around the normal vector:

$$E_k(\alpha_j, \beta_j, \gamma_j) = \rho'_{jk} \int_{\Omega'_i} L_k \left(R_{\alpha_j,\beta_j,\gamma_j}(\theta'_i, \phi'_i) \right) A(\theta'_i) d\Omega'_i \ . \tag{5.4}$$

Then, if the rotation $(R_{\alpha_j,\beta_j,\gamma_j}(\theta'_i, \phi'_i))$ is represented by the rotation matrix R and the local incidence angles are stored in the vector ω'_i the equation (5.4) may be written as:

$$E_k(\mathsf{R}) = \rho'_{jk} \int_{\Omega'_i} L_k \left(\mathsf{R}\omega'_i \right) A(\omega'_i) d\omega'_i \ , \tag{5.5}$$

where this result is analogous to the spatial convolution of the functions f and g given by:

$$(f \otimes g)(a) = \int_a f(T_a(t)) g(t) dt \tag{5.6}$$

In the case of a linear convolution, $T_a(t)$ corresponds to a translation of $(t - a)$. If T_a is a rotation by the angle a, then formula (5.6) defines the convolution in the angular domain. Comparing the equations (5.6) and (5.5) we can see many similarities and it is possible to consider the image irradiance as the result of convolving the incident illumination L by a transfer function $A = \cos\theta'_i$. In brief, the irradiances E_k correspond to different rotations of the incident light field. Additionally, the equation (5.4) can be simplified by transforming the integral into a product of the coefficients describing the spherical harmonics components of the light.

5.1.2 Spherical Harmonics

The spherical harmonics representation is a mathematical tool widely used in the signal processing area to model physical phenomena such as those involving planetary movement or quantum mechanics. They conform a set of functions which is equivalent to the basis describing the Fourier series but expressed on a 2D space using spherical coordinates. Formally, this basis is the solution of the *spherical harmonic differential equation*, which is given by the angular part of Laplace's equation [Groemer 1996].

The spherical harmonics of order $l \geq 0$ and degree m (with $-l \leq m \leq l$) are given by:

$$Y_{lm}(\theta, \phi) = N_{lm} P_{lm}(\cos\theta) e^{Im\phi} \text{ with } I = \sqrt{-1} \ , \tag{5.7}$$

where,

$$N_{lm} = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} \ . \tag{5.8}$$

In equation (5.7) the dependence on the colatitudinal variable (θ) is expressed in terms of the associated Legendre polynomials P_{lm}. These functions are the canonical solution of the general Legendre equation, a second degree differential equation commonly used for modeling physical phenomena. The P_{lm} is expressed in the form:

$$P_{lm}(x) = \frac{(-1)^m}{2^l l!} (1-x^2)^{m/2} \frac{\partial^{l+m}}{\partial x^{l+m}} (x^2-1)^l \ . \tag{5.9}$$

The azimuthal dependence (ϕ) is expanded using the Fourier basis functions. The three first orders of the spherical harmonics $l \in \{0, 1, 2\}$ (only the terms with $m \geq 0$) are shown below. In figure 5.1 we also show the absolute value of these terms.

$$\begin{aligned}
&Y_{00} = \sqrt{\tfrac{1}{4\pi}} \\
&Y_{10} = \sqrt{\tfrac{3}{4\pi}} \cos\theta \qquad\qquad Y_{11} = -\sqrt{\tfrac{3}{8\pi}} \sin\theta e^{I\phi} \\
&Y_{20} = \tfrac{1}{2}\sqrt{\tfrac{5}{4\pi}} (3\cos^2\theta - 1) \quad Y_{21} = -\sqrt{\tfrac{15}{8\pi}} \sin\theta\cos\theta e^{I\phi} \quad Y_{22} = \tfrac{1}{2}\sqrt{\tfrac{15}{8\pi}} \sin^2\theta e^{I2\phi}
\end{aligned} \tag{5.10}$$

Spherical harmonics have an important property which makes them suitable for modeling the illumination. Similar to the case of the Fourier Series, the functions resulting from the spherical harmonics form an orthogonal basis. Thus, any function on the unit sphere can be represented by some coefficients and this basis. In mathematical form, the function $f(\theta, \phi)$ is described by the coefficients f_{lm} in such a way:

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} f_{lm} Y_{lm}(\theta, \phi) \tag{5.11}$$

with,

$$f_{lm} = \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} f(\theta, \phi) Y_{lm}^*(\theta, \phi) \sin\theta \, d\theta d\phi \ .$$

In their work, Ramamoorthi *et al.* developed a rotation operator suitable for the space spanned by the spherical harmonics functions. It is noted by $D^l_{mm'}(\alpha, \beta, \gamma)$. This operator

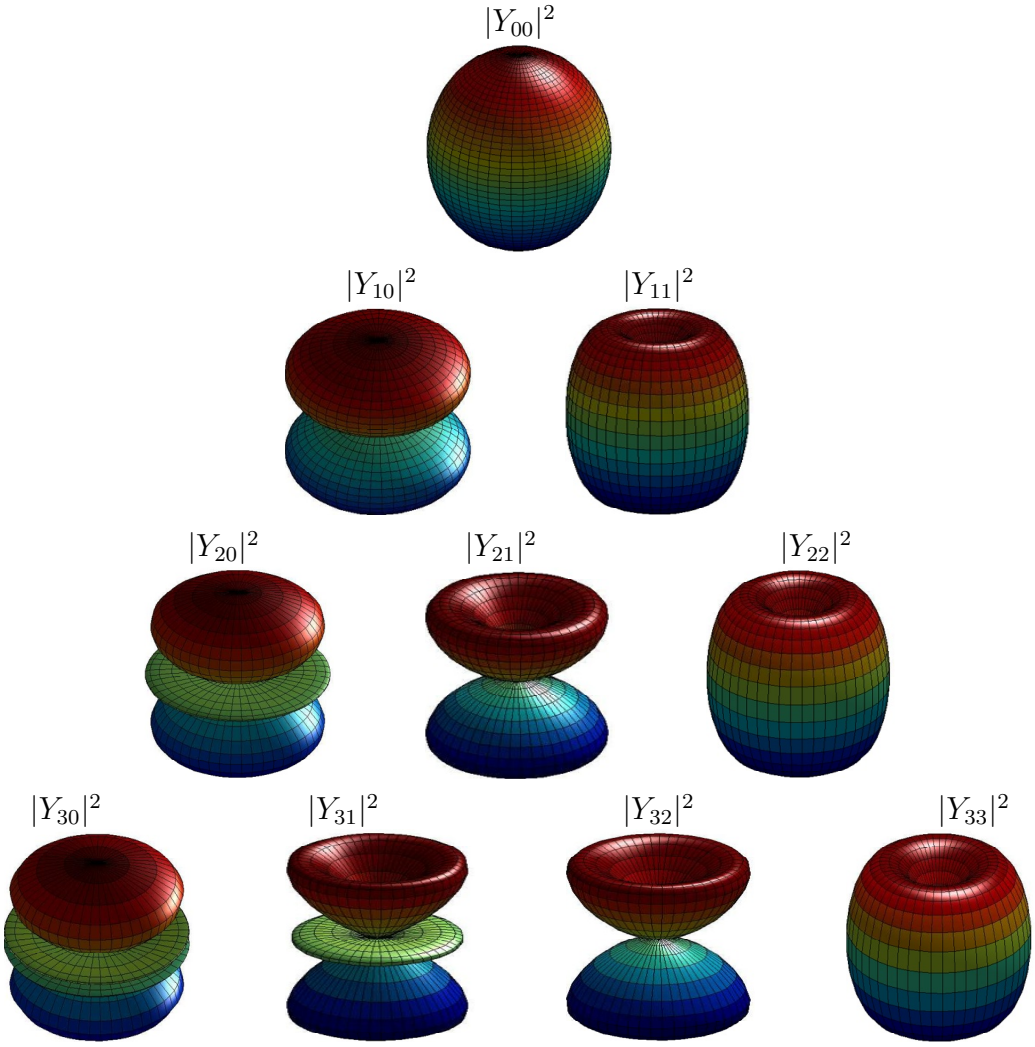


Figure 5.1: Spherical Harmonics for the first 3 orders (only the terms with $m \geq 0$ are shown).

represents a rotation around the normal vector parameterized with (α, β, γ). Then, the rotated basis is rewritten as :

$$Y_{lm}(R_{\alpha,\beta,\gamma}(\theta,\phi)) = \sum_{m'=-l}^{l} D^l_{mm'}(\alpha,\beta,\gamma) Y_{lm'}(\theta,\phi) \tag{5.12}$$

Further, the only term meaningful for the formulation of the reflectance equation is D^l_{m0} which is only dependent on the order and the basis expressed by the rotation angles (α, β):

$$D^l_{m0}(\alpha,\beta,\gamma) = \sqrt{\frac{4\pi}{2l+1}} Y_{lm}(\alpha,\beta) \ .$$

As we will see in the next subsection, this result is extremely practical for the manipulation of equation (5.4). A deeper analysis of the rotation operator can be found in Ramamoorthi's work [Ramamoorthi 2001b].

5.1.3 The Image Formation in Terms of Spherical Harmonics

Using the tools previously described, the components of equation (5.4) can be expressed in the spherical harmonics space. For the illumination, the equation in global coordinates is:

$$L(\theta_i,\phi_i) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} L_{lm} Y_{lm}(\theta_i,\phi_i) \ . \tag{5.13}$$

Using the rotation operator, the illumination in local coordinates is expressed as:

$$L(\theta_i,\phi_i) = L(R_{\alpha,\beta,\gamma}(\theta'_i,\phi'_i)) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sum_{m'=-l}^{l} L_{lm} D^l_{mm'} Y_{lm'}(\theta'_i,\phi'_i) \ . \tag{5.14}$$

On the other hand, the cosines term in the image formation equation is also written in the same basis. It is important to notice that in this case only the upper hemisphere is visible, thus $A(\theta') = 0$ for the invisible lower hemisphere. Also, this term is independent from the azimuthal angle and coefficients A_{ln} with $n \neq 0$ disappear.

$$A(\theta'_i) = \max(\cos\theta'_i, 0) = \sum_{l=0}^{\infty} A_l Y_{l0}(\theta'_i) \ . \tag{5.15}$$

Finally if we multiply the expansions for the illumination and the Lambertian term and integrate over the hemisphere, the equation (5.4) is expressed in terms also of the spherical harmonics basis:

$$E_k(\alpha_j,\beta_j,\gamma_j) = \rho'_{jk} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_l L_{lm} D^l_{m0}(\alpha_j,\beta_j,\gamma_j) Y_{lm}(\alpha_j,\beta_j) \tag{5.16}$$

$$E_k(\alpha_j,\beta_j) = \rho'_{jk} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sqrt{\frac{4\pi}{2l+1}} A_l L_{lm} Y_{lm}(\alpha_j,\beta_j) \ . \tag{5.17}$$

We can also expand the image irradiance for the point $\mathbf{x}$ using the spherical harmonics decomposition:

$$E(\alpha_j, \beta_j) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} E_{lm}^{j} Y_{lm}(\alpha_j, \beta_j) \ . \tag{5.18}$$

Comparing (5.17) and (5.18) shows that the coefficients representing the irradiance E_{lm} and the coefficients of the illumination and the cosines term are closely related:

$$E_{lm}^{j} = \rho'_{jk} \sqrt{\frac{4\pi}{(2l+1)}} A_l L_{lm} \ . \tag{5.19}$$

This result justifies the theory proposed by Ramamoorthi *et al.* : the integral over the hemisphere representing the interaction of the illumination and the surface reflectance properties is transformed into a simple multiplication of coefficients in the frequency framework established by the spherical harmonics basis. However we still have to define the values A_l. They correspond to the spherical harmonics coefficients of the cosines function. If we insert the transfer function $A(\theta'_i)$ in equation (5.8), the coefficients are given by:

$$A_l = 2\pi \int_0^{\frac{pi}{2}} Y_{l0}(\theta'_i) \cos\theta'_i \, \sin\theta'_i \, d\theta'_i \tag{5.20}$$

The results of this integral are constant values dependent on l. Then, after including the square root term, we can represent equation (5.19) as a matrix product, here denoting only the sixth first orders for the spherical harmonics:

$$\mathbf{E}_j = \rho'_{jk} \, \mathrm{diag}(\mathbf{A}) \mathbf{L} \ , \tag{5.21}$$

where $\mathbf{E}_j = \begin{bmatrix} E_{00}^{j} & E_{1m}^{j} & E_{2m}^{j} & \cdots & E_{6m}^{j} \end{bmatrix}^{\mathsf{T}}$, $\mathbf{L} = \begin{bmatrix} L_{00} & L_{1m} & L_{2m} & \cdots & L_{6m} \end{bmatrix}^{\mathsf{T}}$ and $\mathrm{diag}(\mathbf{A})$ corresponds to a diagonal matrix whose nonzero elements are:

$$\left\{ \pi, \ \ \frac{2\pi}{4}, \ \ \frac{\pi}{4}, \ \ 0, \ \ -\frac{\pi}{24}, \ \ 0, \ \ \frac{\pi}{64} \right\}$$

The linear representation of this interaction opens the door to a simple but effective modeling of the image irradiance in terms of a basis and the coefficients describing a function on the unit sphere. The most relevant result for our work of the analysis described in previous paragraphs is that the coefficients $\mathbf{E}_j$ describing the irradiance of a point $\mathbf{x}$ can be expressed approximately using an order 2 of the spherical harmonics. Since there are $2l + 1$ terms for

each order, we need in total 9 terms to model the irradiance of a Lambertian surface.

$$\begin{bmatrix} E^j_{0,0} \\ E^j_{1,-1} \\ E^j_{1,0} \\ E^j_{1,1} \\ E^j_{2,-2} \\ E^j_{2,-1} \\ E^j_{2,0} \\ E^j_{2,1} \\ E^j_{2,2} \end{bmatrix} = \rho'_{jk} \begin{bmatrix} \pi & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2\pi}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{2\pi}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{2\pi}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\pi}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\pi}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\pi}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\pi}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\pi}{4} \end{bmatrix} \begin{bmatrix} L_{0,0} \\ L_{1,-1} \\ L_{1,0} \\ L_{1,1} \\ L_{2,-2} \\ L_{2,-1} \\ L_{2,0} \\ L_{2,1} \\ L_{2,2} \end{bmatrix} \tag{5.22}$$

$$\mathbf{E}_j = \rho'_{jk} \mathsf{C} \mathbf{L} \ , \tag{5.23}$$

where C is the sparse matrix included in (5.22).

5.2 Non–linear Optimisation

We have now a function to express the irradiance of a point of the surface as a linear combination of the spherical harmonics coefficients of the illumination and the surface reflectance. Similar to the procedure described in 4.2, we can formulate a non linear system in order to estimate the unknown parameters. Next, we will explain in detail some of the parts involved on this process.

5.2.1 Description of the minimization function

We denote the estimated value of the image intensity for a surface element j projected in camera i by $\hat{B}_{ijk}$. Let us recall that in traditional cameras there exist three intensity values for the projection of a surface element, represented by the subindex $k \in \{R, G, B\}$. This is our case, however the procedure can be easily generalized to more or less color channels.

In the previous section we explained that the irradiance can be successfully approximated from the spherical harmonics analysis using coefficients of order $l \leq 2$. Once this is done, one can apply Grossberg and Nayar's model for the camera response function [Grossberg 2004] to predict the intensity values:

$$\hat{B}_{ijk} = f_{ik} \left(\sum_{l=0}^{2} \sum_{m=-l}^{l} E^{ijk}_{lm} Y_{lm}(\alpha_j, \beta_j) \right) \ , \tag{5.24}$$

where the values E^{ijk}_{lm} are found as shown in equations (5.22) and (5.23). Here, the illumination present in image i is represented by the spherical harmonic coefficients stored in vector ($\mathbf{L}_{ik}$). Then the coefficients E^{ijk}_{lm} are expressed in matrix form as follows:

$$\mathbf{E}_{ijk} = \rho'_{jk} \mathsf{C} \mathbf{L}_{ik} \ . \tag{5.25}$$

Expression (5.24) can also be represented in matrix form. The values for the basis functions $Y_{lm}(\alpha_j, \beta_j)$ presented in equation (5.10) for the point j build up the following vector:

$$\mathbf{Y}_j = \begin{bmatrix} Y^j_{0,0} & Y^j_{1,-1} & Y^j_{1,0} & Y^j_{1,1} & Y^j_{2,-2} & Y^j_{2,-1} & Y^j_{2,0} & Y^j_{2,1} & Y^j_{2,2} \end{bmatrix}^\mathsf{T} ,$$

(here until order $l = 2$). Thus, the image irradiance equation can be formulated as a matrix operation:

$$\hat{B}_{ijk} = f_{ik}\left(\rho'_{jk}(\mathsf{C}\mathbf{L}_{ik})^\mathsf{T}\mathbf{Y}_j\right) \quad . \tag{5.26}$$

Additionally, we already know that $f_{ik}(E)$ can be modeled by a linear combination of coefficients plus a function basis (*cf.* equation (2.9)). Therefore, the final intensity registered in the image is estimated by:

$$\hat{B}_{ijk} = h_0\left(\rho'_{jk}(\mathsf{C}\mathbf{L}_{ik})^\mathsf{T}\mathbf{Y}_j\right) + \sum_{n=1}^{N} w_{ikn} h_n\left(\rho'_{jk}(\mathsf{C}\mathbf{L}_{ik})^\mathsf{T}\mathbf{Y}_j\right) \quad . \tag{5.27}$$

Given the 3D structure of the scene the unknown parameters in this equation are: The albedo ρ'_{jk}, the CRF coefficients w_{ikn} and the spherical harmonics coefficients for the illumination L_{ik}. One can group the per–image terms depending on the subindex i in vector $\mathbf{a}_i = \begin{bmatrix} L^\mathsf{T}_{iR} & L^\mathsf{T}_{iG} & L^\mathsf{T}_{iB} & \mathbf{w}^\mathsf{T}_{iR} & \mathbf{w}^\mathsf{T}_{iG} & \mathbf{w}^\mathsf{T}_{iB} \end{bmatrix}^\mathsf{T}$ where the vector $\mathbf{w}_{ik}$ contains the CRF coefficients ($w_{ikn}, \forall n = 1 \cdots N$). If we use O spherical harmonics coefficients for illumination and N coefficients for the CRF the dimension of $\mathbf{a}_i$ is $(3 \times O) + (3 \times N)$. In turn, the vector $\mathbf{b}_j$ of dimension 3 represents a variable proportional to the surface albedo: $\mathbf{b}_j = \begin{bmatrix} \rho'_{jR} & \rho'_{jG} & \rho'_{jB} \end{bmatrix}^\mathsf{T}$ (remember that $\rho'_{jk} = (\rho_j \cdot q_k)/\pi$).

In equation (4.14), we formulated our parameter estimation in the framework of a minimization problem. This time, we will proceed similarly, with the difference that the estimated value of the brightness is parameterized by the two mentioned vectors $(\mathbf{a}_i, \mathbf{b}_j)$. Thus, the estimation function $\boldsymbol{B}(\mathbf{a}_i, \mathbf{b}_j)$ is defined as a function from $\Re^{3\times(O+N+1)} \rightarrow \Re^3$.To estimate the unknowns $\mathbf{a}_i$, $\mathbf{b}_j$, we minimize the difference between the observed and predicted intensity values. An optimal solution to calculate the unknowns requires a full non-linear optimization of the cost function, defined as the squared difference between the measured intensity and its corresponding estimation. Given a set of J surface elements projected in M images, the optimization problem is formulated as follows: for M images composing the set, and a 3D model formed by J surface elements we minimize the following function:

$$\underset{\mathbf{a}_i, \mathbf{b}_j}{\operatorname{argmin}} \sum_{i=1}^{M}\sum_{j=1}^{J} \left(\mathbf{B}_{ij} - v_{ij}\boldsymbol{B}(\mathbf{a}_i, \mathbf{b}_j)\right)^2 \quad . \tag{5.28}$$

Similar to the case presented in previous chapter, v_{ij} models the visibility of a point j in a camera i. Also, note that unknowns can not be estimated without ambiguity: albedos ρ'_j and illumination coefficients $\mathbf{L}_i$ can only be computed up to one global scale factor. As in the non linear minimization case illustrated in section 4.2 we impose a monotonocity constraint

on the coefficients of the CRF's: for 1024 equidistant irradiance values between 0 and 1, we enforce that the CRF is larger for each sampled irradiance value than for the next smaller one. An important note in this process is that we are implicitly assuming that camera sensitivity $Q_k(\lambda)$ described in (2.1) is similar for all images.

5.2.2 Sparse Photometric Adjustment: a parallel to Sparse Bundle Adjustment

In section 2.1.1 we had shortly described the problem of geometric multi–view reconstruction and particularly, we mentioned the bundle adjustment algorithm as a standard final step in order to recover the camera parameters and the position of a set of 3D points. This problem can be formulated as follows: given a set of image features represented by their 2D coordinates $\mathbf{x}_{ij}$ the goal is to find the set of camera matrices P_i and the 3D points $\mathbf{X}_j$ such that $\mathbf{x}_{ij} = \mathsf{P}_i\mathbf{X}_j$. As frequently happens in real world scenarios, measurements of the 2D points are noisy and the relation $\mathbf{x}_{ij} = \mathsf{P}_i\mathbf{X}_j$ can not be satisfied exactly. In this case a non–linear estimation is proposed in such a way that the solution satisfies the Maximum Likelihood of the cost function under the assumption of a Gaussian noise. This solution is found by minimizing the Euclidean distance between the position calculated by the estimation function $f(\mathsf{P}_i, \mathbf{X}_j)$ and the values measured by the samples $\mathbf{x}_{ij}$. Then, for J 3D points projected in M images we have to solve:

$$\underset{\mathsf{P}_i,\mathbf{X}_j}{\operatorname{argmin}} \sum_{i=1}^{M}\sum_{j=1}^{J} |\mathbf{x}_{ij} - f(\mathsf{P}_i, \mathbf{X}_j)|^2 \quad , \tag{5.29}$$

where $|\cdot|^2$ denotes the Euclidean distance.

The similarity between equation (5.28) and equation (5.29) is evident: if we focus on the estimation function, both of them depend on parameters associated to the camera (subindex i) and the 3D point (subindex j). This particularity was first mentioned in [Luong 2002], where authors try to recover a simple linear camera response function along with light directions and albedos. The resemblance is also manifest when using a generic camera model for general formulation of the bundle adjustment problem like authors do in [Lourakis 2009].

But let us explore slightly deeper this relationship. In our case, our estimation function $B(\mathbf{a}_i, \mathbf{b}_j)$ finds a triplet of computed intensities based on the vector $\mathbf{a}_i$ that groups the camera and illumination parameters and the vector $\mathbf{b}_j$ modeling the point surface reflectance (albedos). In a typical bundle adjustment problem, we aim to find for each camera a projection matrix P_i and the position of the set of 3D points $\mathbf{X}_j$; the projection matrix is proportional to the camera intrinsic parameters and its pose ($\mathsf{P}_i \sim \mathsf{K}_i\mathsf{R}_i\begin{pmatrix}\mathbf{I} & -\mathbf{t}_i\end{pmatrix}$) and sometimes a geometric distortion model is also included in this computation.

At this point, the similarities are more than evident:

- While for the bundle adjustment algorithm we have measurements for the positions of

the points on the image ($\mathbf{x}_{ij}$) in our formulation we have the intensity values in the images ($\mathbf{B}_{ij}$).

- In the geometric reconstruction problem we estimate the 3D point positions ($\mathbf{X}_j$) while in the appearance estimation problem we find the albedos ($\mathbf{b}_j$) for a surface element on the 3D model.
- In the bundle adjustment problem we obtain for each camera a projection matrix P_i along with its intrinsic parameters K_i. In our approach we get for each image an approximation of the reflected illumination $\mathbf{L}_{ik}$ and the coefficients for the camera response functions $\mathbf{w}_{ik}$.
- The classical bundle adjustment process is based on a pinhole camera model (however other camera models can be used). Our method uses a Lambertian model for the surface reflectance properties and the illumination represented as spherical harmonics.
- The main constraint in classical bundle adjustment allowing the formulation of the projective reconstruction is the rigidity of the scene [1]. In turn our formulation assumes that the albedo for a surface element is unique.

Inspired by the table presented in [Luong 2002], we introduce our version of the previously described comparison in table 5.2.2. It summarizes the parallelism of the properties present in the bundle adjustment problem and the solution here proposed.

5.2.3 Robust Estimation

Solving the optimization problem described in section 5.2.1 is not a trivial task. Fortunately efficient methods, like those presented in the section 2.3, can be applied to our formulation. The Levenberg-Marquardt optimization aims at finding the Maximum Likelihood solution for the equation (5.28) under the assumption of Gaussian noise. This assumption is made implicitly when we choose the Euclidean distance between the samples and the estimated value as criterion to minimize. However is some cases this criterion might not be the most appropriate. The selection of the cost function used in the minimization will influence the efficiency of computation and mainly the accuracy of the estimated parameters. Also, the capacity to manage errors in the measurements is highly engaged. A pedagogical example of the importance of selecting a coherent criterion is presented in [Zhang 1997], using a conic fitting problem to illustrate the process.

But, particularly speaking, why is it important to use a robust estimation in our case? Because the proposed pipeline to compute jointly the photometric properties of the scene (Illumination, surface reflectance and CRFs) presents a considerably large number of possible sources of error. The first one is the quality of the 3D model. Even if the number of

[1] There exist particular adaptations of bundle adjustment where the scene is not rigide, for example, when working with deformable objects as the human body or animals.

	Bundle adjustment problem	**Photometric joint estimation**
Measures	Image points: $\mathbf{x}_{ij}$	Intensities: $\mathbf{B}_{ij}$ 3D structure (normals): $\mathbf{n}_j = [n_{j1}\ n_{j2}\ n_{j3}]^\mathsf{T}$
Unknowns	3D point positions: $\mathbf{X}_j = [X_{1_j}\ X_{2_j}\ X_{3_j}]^\mathsf{T}$ Camera projection matrix: $\mathsf{P}_i \sim \mathsf{K}_i\mathsf{R}_i \begin{pmatrix}\mathbf{I} & -\mathbf{t}_i\end{pmatrix}$	Surface reflectance: (albedo) $\mathbf{b}_j = \begin{bmatrix}\rho'_{jR} & \rho'_{jG} & \rho'_{jB}\end{bmatrix}^\mathsf{T}$ Illumination + CRF coefficients: $\mathbf{a}_i = [L_{iR}^\mathsf{T}\ L_{iG}^\mathsf{T}\ L_{iB}^\mathsf{T}\ \mathbf{w}_{iR}^\mathsf{T}\ \mathbf{w}_{iG}^\mathsf{T}\ \mathbf{w}_{iB}^\mathsf{T}]^\mathsf{T}$
Calibration	Camera intrinsic parameters. K_i	Camera response functions (CRF's) $f(E) = h_0(E) + \sum_{n=1}^{N} w_n h_n(E)$
Constraint	A static scene.	Photometric properties remain unchanged. Distant light source.

Table 5.1: Similarities between the Bundle Adjustment problem and the Photometric joint estimation problem.

possible wrong points issued from the multi–view reconstruction process is limited by refining a mesh, the 3D reconstruction is not perfect. For example, the photographs involved in the reconstruction usually do not enclose all the target objects (*e.g.* the roof of a building, or its posterior facade) and the 3D reconstructed model is an incomplete structure presenting discontinuities where the image information is no available.

A second source of error is the accuracy of the cameras geometric calibration. We use the approach of Snavely *et al.* [Snavely 2008] to recover the camera pose and its intrinsic parameters. This method exhibits an excellent performance when working with real images. However, sporadically, bad estimations of the projection matrix for specific images are calculated. We avoid the presence of these issues by removing manually the evidently problematic images (those ones were the projection of the mesh on the 2D plane do not correspond with the original image). Also when the images contain objects not modeled by the 3D structure (*e.g.* pedestrians, trees, cars, etc.) the projection of the 3D point on the image(s) could yield to the wrong colors.

Another source of error to work around is the presence of points violating the diffuse reflectance assumption (windows on facades or metallic surfaces). If one of the mentioned cases is present during the minimization of the function (5.28), then the difference between the measured value and the estimated intensity should be large. A common method robust to the samples presenting this behavior is formulated by using the **Iterative Reweighted Least Squares** (IRLS) algorithm. The principle of the method is to assign a weight to the

measurements, depending on the residual error issued from the preceding iteration. Let us denote the residual error in our minimization as:

$$r_k(\mathbf{a}_i, \mathbf{b}_j) = \mathbf{B}_{ij} - v_{ij}\boldsymbol{B}(\mathbf{a}_i, \mathbf{b}_j) \ . \tag{5.30}$$

In the IRLS scheme, the problem can be formulated as follows:

$$\min_{\mathbf{a}_i, \mathbf{b}_j} \sum_{k=1}^{K} u\left(r_k^{\text{it}-1}\right) r_k^2 \ . \tag{5.31}$$

The problem is solved by a general Least Squares minimization, *e.g.* using Levenveng-Marquardt, but instead of having an L_2 cost function the values of the residuals at each iteration are multiplied by the weight $u\left(r_k^{\text{it}-1}\right)$ computed at the last iteration. There exist a large variety on the forms for the function u. In statistics literature estimation methods based on such functions are also known as M-estimators. There is no consensus about which function can perform the best results, however some insigths can be obtained by exploring the nature of the noise. In our work we have tested two functions (*Cauchy* and *Huber*) and the final implementation employs the one showing the more esthetically pleasant results (*Huber*) (differences are almost imperceptible).

In table 5.2 we show the analytical expressions for cost function $f(r)$ and weight functions $u(r)$ and their corresponding graphics in the case of the Euclidean distance (L_2) (used in traditional Least Squares problems), Huber and Cauchy M-estimators used by our robust estimation.

For the evaluation of the new robust cost function $u(r)$, it is necessary to define the parameter that rules the tolerance to outliers during the estimation. For the case of the Cauchy function this parameter is represented by c, for the Huber function is called k. An extensive discussion on how to select these tuning parameters according to the variance on the samples is presented in [Zhang 1997].

5.2.4 Jacobian matrix structure

One particular advantage from the formulation of $\boldsymbol{B}(\mathbf{a}_i, \mathbf{b}_j)$ in the equation (5.28) as a function depending on the vectors $\mathbf{a}_i$ and $\mathbf{b}_j$ is that the Jacobian matrix can be easily expressed as a sparse structure. From the practical point of view, this representation has many benefits. The processing time is reduced substantially as well as the memory usage. Since we are performing a joint estimation of the parameters, where the number of unknowns to estimate could easily exceed some thousands of values (using a hundred of images), the fact of expressing the Jacobian matrix in this way is very convenient.

The main observation for arranging the elements of the Jacobian matrix is that all the parameters of the different cameras do not interact at the same time, when evaluating the cost function of equation (5.31) for a particular measure. This characteristic is also common with the Sparse Bundle Adjustment algorithm, although it was not mentioned in the

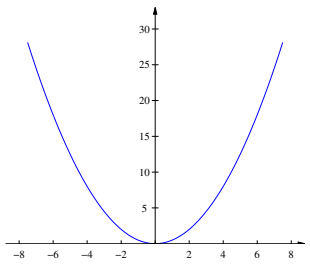
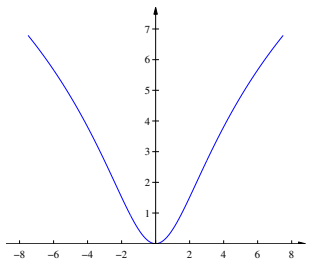
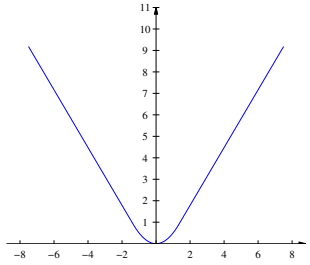
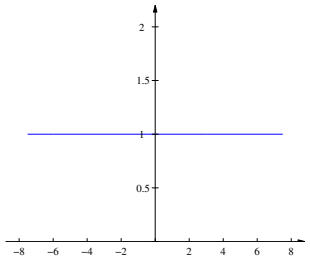
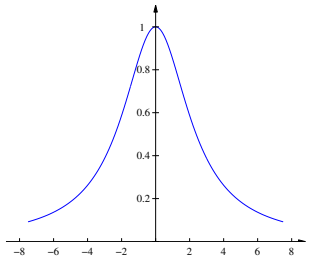
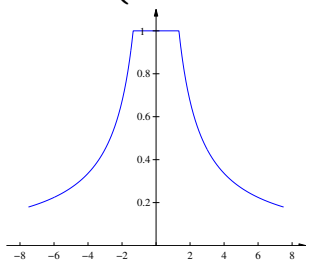
	L_2	Cauchy	Huber $\begin{cases}\text{if } \lvert r\rvert \geq k \\ \text{if } \lvert r\rvert < k\end{cases}$
$f(r)$	$\frac{r^2}{2}$ 	$\frac{c^2}{2}\log\left(1+(r/c)^2\right)$ 	$\begin{cases} r^2/2 \\ k(\lvert r\rvert - k/2)\end{cases}$ 
$u(r)$	1 	$\frac{1}{1+(r/c)^2}$ 	$\begin{cases} 1 \\ k/\lvert r\rvert\end{cases}$ 

Table 5.2: Cost functions used for the robust estimation.

comparison presented in section 5.2.2. In this context, the structure of the Jacobian matrix is similar to that introduced for the case of sparse bundle adjustment (see Appendix 6 of [Hartley 2003] and [Lourakis 2009]). Formally, the Jacobian matrix is found by computing the partial derivatives of the residual error $r_{ij}(\mathbf{x}) = \mathbf{B}_{ij} - v_{ij}\boldsymbol{B}(\mathbf{a}_i, \mathbf{b}_j)$, where the vector $\mathbf{x} = [\mathbf{a}_1^\mathsf{T}\ \mathbf{a}_2^\mathsf{T}\ \cdots\ \mathbf{a}_M^\mathsf{T}\ \mathbf{b}_1^\mathsf{T}\ \mathbf{b}_2^\mathsf{T}\ \cdots\ \mathbf{b}_J^\mathsf{T}]^\mathsf{T}$. For simplicity we suppose here that all the points are visible in all the images: there are $Q = 3 \times (J \times M)$ measurements. Also, as mentioned before, the number of parameters to estimate is $P = 3 \times (O + N + 1)$ (*cf.* section 5.2.1).

Hence, the Jacobian matrix is:

$$\mathsf{J}(\mathbf{x}) = \begin{pmatrix} \frac{\partial r_{11}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{11}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{11}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{11}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \frac{\partial r_{12}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{12}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{12}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{12}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_{1J}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{1J}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{1J}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{1J}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \frac{\partial r_{21}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{21}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{21}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{21}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \frac{\partial r_{22}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{22}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{22}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{22}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_{2J}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{2J}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{2J}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{2J}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_{M1}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{M1}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{M1}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{M1}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \frac{\partial r_{M2}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{M2}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{M2}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{M2}}{\partial \mathbf{b}_J}(\mathbf{x}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_{MJ}}{\partial \mathbf{a}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{MJ}}{\partial \mathbf{a}_M}(\mathbf{x}) & \frac{\partial r_{MJ}}{\partial \mathbf{b}_1}(\mathbf{x}) & \cdots & \frac{\partial r_{MJ}}{\partial \mathbf{b}_J}(\mathbf{x}) \end{pmatrix}_{[Q \times P]} \tag{5.32}$$

Let us examine a little closer the structure of this matrix. For this, we will take as example a matrix with $J = 4$ points projected on $M = 3$ images. We get two block of well defined columns: those ones with partial derivatives related to per–image parameters ($\mathbf{a}_i$) and the columns with partial derivatives with respect to the surface element parameters $\mathbf{b}_j$. On the rows of the matrix we get also blocks denoting the residual error for a all the points j in each image i. The key observation is that the terms $\frac{\partial r_{ij}}{\partial \mathbf{a}_k} = 0$ and $\frac{\partial r_{ij}}{\partial \mathbf{b}_k} = 0$, $\forall\ i \neq k$, $\forall\ j \neq k$.

Hence, the Jacobian matrix for our example becomes:

$$\mathsf{J}(\mathbf{x}) = \begin{pmatrix} \frac{\partial r_{11}}{\partial \mathbf{a}_1}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \frac{\partial r_{11}}{\partial \mathbf{b}_1}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{\partial r_{12}}{\partial \mathbf{a}_1}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial r_{12}}{\partial \mathbf{b}_2}(\mathbf{x}) & \mathbf{0} & \mathbf{0} \\ \frac{\partial r_{13}}{\partial \mathbf{a}_1}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial r_{13}}{\partial \mathbf{b}_3}(\mathbf{x}) & \mathbf{0} \\ \frac{\partial r_{14}}{\partial \mathbf{a}_1}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial r_{14}}{\partial \mathbf{b}_4}(\mathbf{x}) \\ \mathbf{0} & \frac{\partial r_{21}}{\partial \mathbf{a}_2}(\mathbf{x}) & \mathbf{0} & \frac{\partial r_{21}}{\partial \mathbf{b}_1}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial r_{22}}{\partial \mathbf{a}_2}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \frac{\partial r_{22}}{\partial \mathbf{b}_2}(\mathbf{x}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial r_{23}}{\partial \mathbf{a}_2}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial r_{23}}{\partial \mathbf{b}_3}(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \frac{\partial r_{24}}{\partial \mathbf{a}_2}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial r_{24}}{\partial \mathbf{b}_4}(\mathbf{x}) \\ \mathbf{0} & \mathbf{0} & \frac{\partial r_{31}}{\partial \mathbf{a}_3}(\mathbf{x}) & \frac{\partial r_{31}}{\partial \mathbf{b}_1}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{\partial r_{32}}{\partial \mathbf{a}_3}(\mathbf{x}) & \mathbf{0} & \frac{\partial r_{32}}{\partial \mathbf{b}_2}(\mathbf{x}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{\partial r_{33}}{\partial \mathbf{a}_3}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \frac{\partial r_{33}}{\partial \mathbf{b}_3}(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{\partial r_{34}}{\partial \mathbf{a}_3}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial r_{34}}{\partial \mathbf{b}_4}(\mathbf{x}) \end{pmatrix} \tag{5.33}$$

The scheme in figure 5.2 illustrates clearly this structure for the example described before, where darker regions represent zeros. The implementation of the optimization was done by adapting the software package **sba** [Lourakis 2009] to our formulation. This software proposes solutions for the bundle adjustment problem based on the sparse Levenberg–Marquardt algorithm. It has great versatility because any generic camera model can be integrated: the user can directly create the function to minimize. In our case, the "camera model" is in fact the image formation model described by $\boldsymbol{B}(\mathbf{a}_i, \mathbf{b}_j)$. The implementation also takes full advantage of the sparsity nature of the problem.

As in the homologue case discussed in chapter 4, we also create a `Matlab` version of the non–linear minimization, this time using the function `lsqnonlin`. It is worth mentioning that in both cases, the space of searches during the non–linear minimization was constrained by boundary conditions depending on the nature of the parameters: for the albedo $\rho \in [0, 1]$, for the CRF coefficients $\mathbf{w}_i \in [w_{\text{min}}, w_{\text{max}}]$ where w_{min} is the minimal coefficient found when representing the set of 201 real world CRF's by the model proposed by Grossberg, and the corresponding maximum value for w_{max}. The spherical harmonics coefficients for the illumination were not bounded.

5.3 Results

We evaluate the performance of our algorithm in real world conditions using the two databases described in section 4.3.2. Both collections target architectural structures in outdoor environments. The first dataset, called DB1, is composed by the 120 images acquired on our own

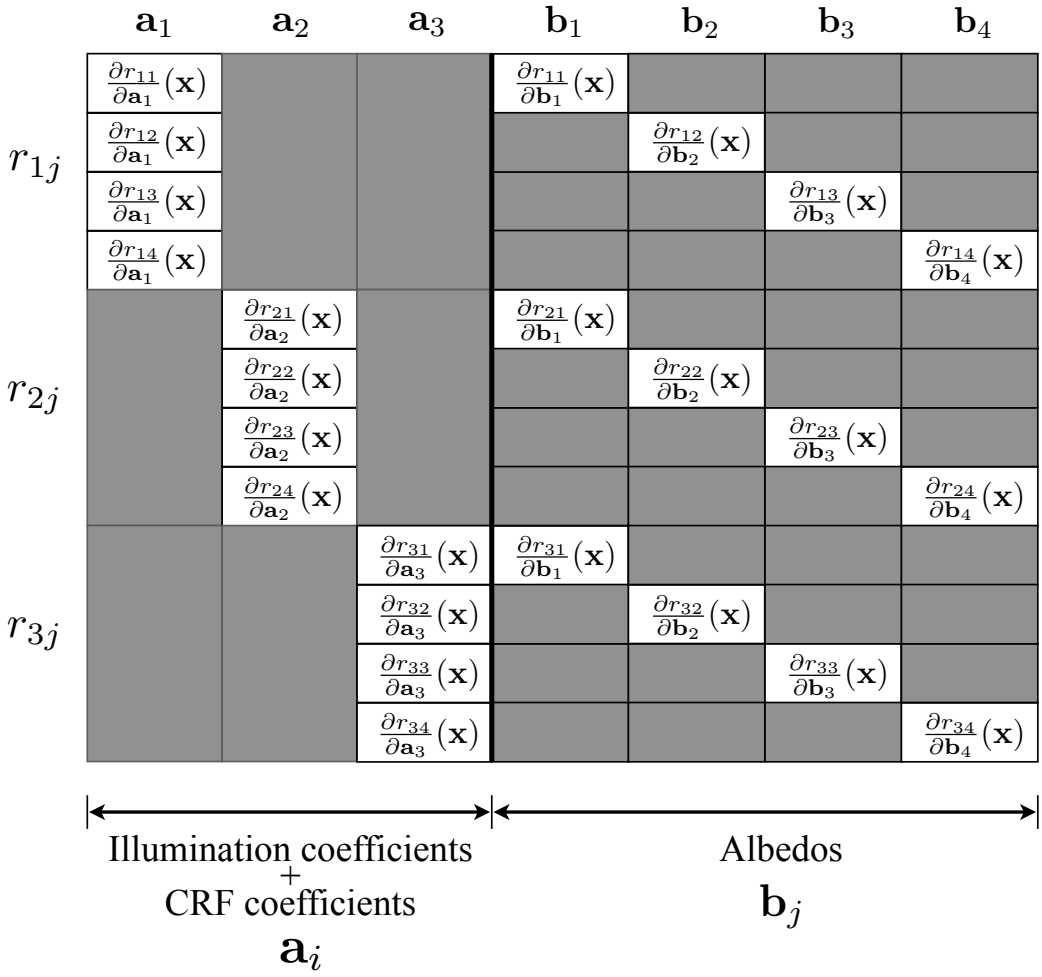


Figure 5.2: Structure of the Jacobian matrix for the minimization function (5.28) when the number of surface elements is $J = 4$ and the number of images is $M = 3$.

and a 3D structure represented by a mesh with 112,504 vertices. We got 10 extra images containing a color checker board inside the scene; we also took multiple exposure images for these extra samples, just seconds after the image used for reconstruction was taken. The second set of data (DB2) is composed by the images of the *Sacré Coœur's* Cathedral collected from an internet image repository and a mesh with 80,444 vertices created from more than nine hundred photos.

To solve our estimation problem, we modeled the CRF with 3 coefficients ($N = 3$) and we used 9 spherical harmonics coefficients to model the illumination ($O = 9$). Let us recall that the number of parameters to estimate is $3 \times (J + M \times (O + N))$, where J is the number of surface elements on the mesh and M the number of images. We divided the first dataset into two subgroups. We estimated the parameters first using 60 images, then for the other 60 using the same mesh. The optimization problem on the first and second subset

used 1,544,338 and 1,622,680 color samples respectively to estimate 339,672 parameters. Using the second database, we limited the estimation to 103 images, with 1,861,539 color samples to estimate 244,932 parameters. These magnitudes testify the large scale nature of the optimization problem.

We focused our evaluation mainly on the estimation of the parameters in the vector $\mathbf{b}_i$, that is the vector describing the camera (CRF coefficients) and the illumination (spherical harmonic coefficients). At the moment it is not evident how to obtain a ground truth for the surface reflectance properties of an outdoor scene, without using expensive and sophisticated equipments (like for example gonioreflectometers). However we present some insights on the stability of our algorithm with respect to the estimation of the albedo. For example, figure 5.4 shows the albedo estimated for the same structure using the two subgroups of the first dataset. The first two columns of the figure show the structure filled up with the calculated albedos. The third column presents the structure under the same point of view where darker regions exhibit large differences between the estimated values of the surface reflectance using the two subsets. The estimated albedo in both cases is very similar for large regions on the surface, which gives us an indicator of the stability of our algorithm.

In figure 5.3 the final weights obtained from applying the robust functions on the residuals of the albedos are displayed. Brighter tones represent regions where the difference between the samples and the estimated values is large. This happens especially in the edges of the model, where the normals of the 3D structure seem to be noisy or not accurate at all.

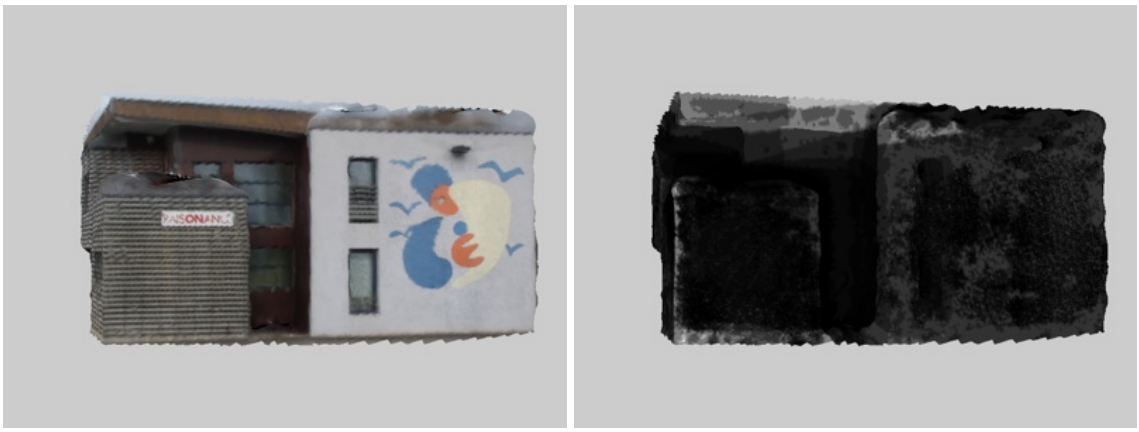


Figure 5.3: Final weights $u\left(r_k^{\text{final}}\right)$ obtained from applying the robust cost function on the error of the estimation function (see equation (5.31)). Brighter values represent large difference between the samples and the estimated values.

5.3.1 CRF Estimation

Since we are using the same set of images used in the result section (4.3) of previous chapter, we have already access to a readily available ground truth for some images of the database.

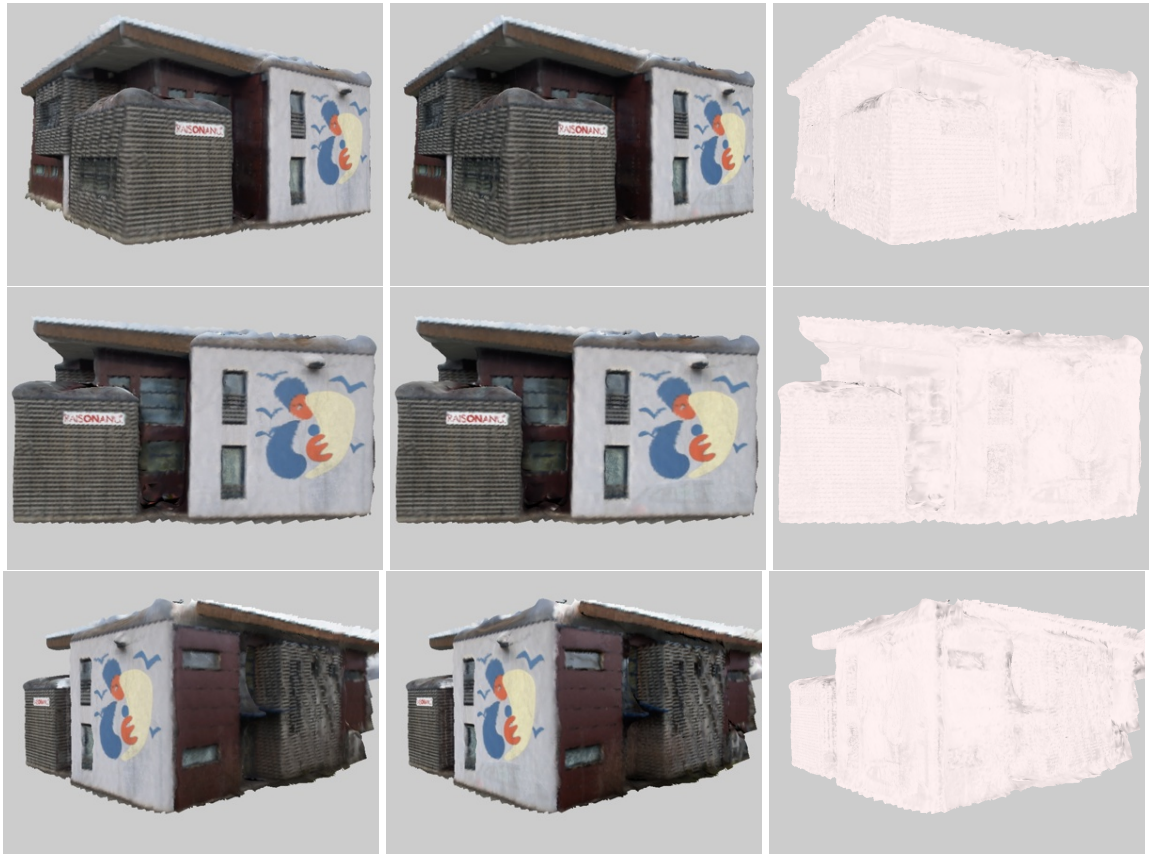


Figure 5.4: Albedo estimated for two subgroups of images of the first dataset. The first two columns show the estimated albedo and the third column presents the difference of the two estimations (darker zones correspond to large differences).

We compare the CRFs estimated by our "complex–illumination" algorithm and the method previously described, where light was represented by a simple model (figure 5.5).

Maybe not surprisingly, we found that the results are very similar to those computed by the "simple–illumination" algorithm. This can be explained because of the strong constraints fixed by the camera response model. The space of possible camera response functions is bounded and the best choices to minimize the optimization error are limited. Also, these results can be justified because the initial points used to start the minimization process are the same in both implementations. The non-linear optimization algorithms are designed to find local optimum points which in the two cases converge to similar results, at least for the case of the CRF coefficients.

5.3.2 Illumination Estimation

A simple approach to evaluate the performance of our technique when estimating the illumination of the scene, consists in calculating rendered images using the approximation of

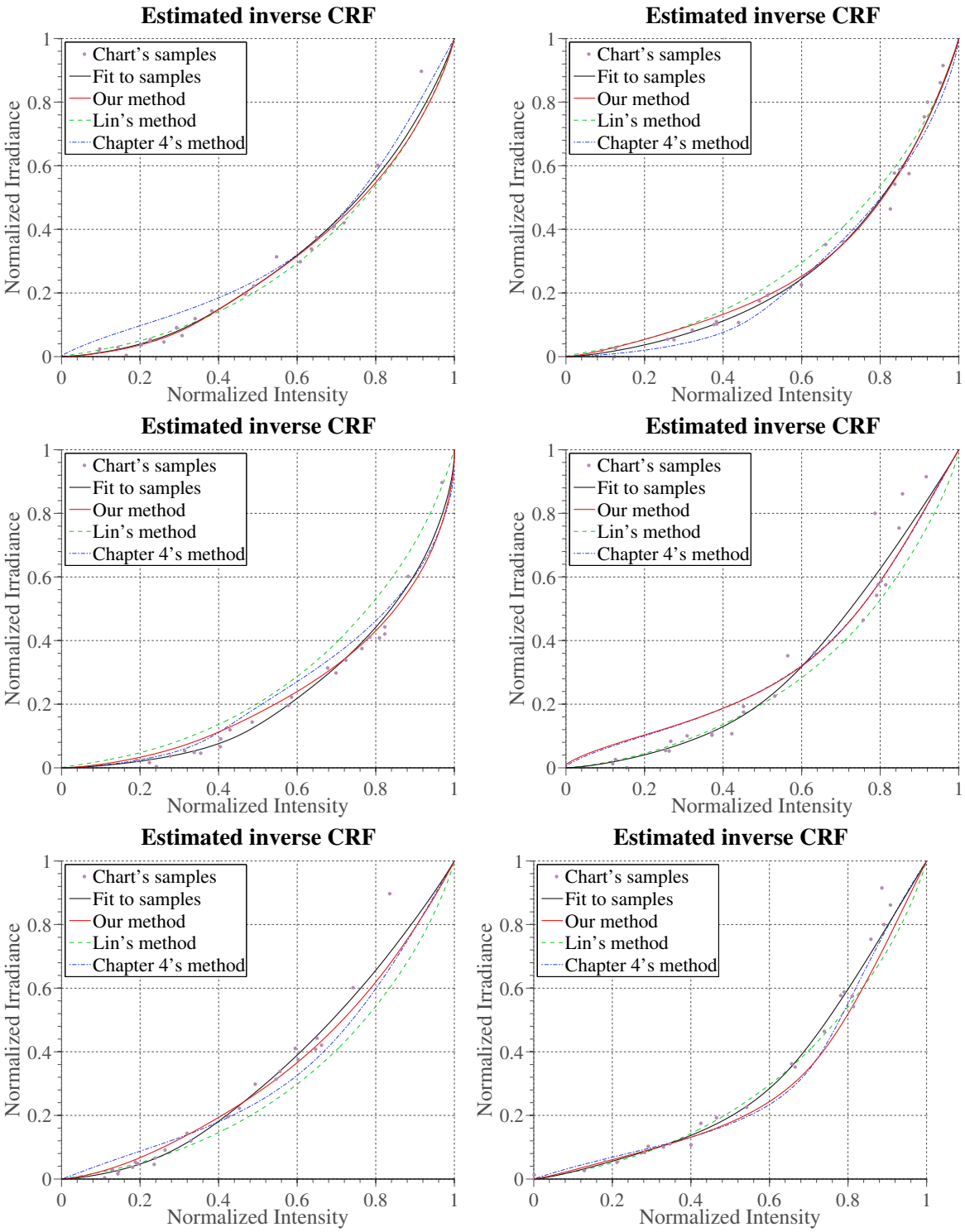


Figure 5.5: CRF estimation using complex illumination model

the lighting found by our algorithm (cross–validation test). For this, we run the process over one of two subsets (*i.e.* the *training* subset) and we extract the estimation for the albedo. Then, we render synthesized versions of the images belonging to a second subset (*test* subset) with the CRF and the spherical illumination calculated on the same group, but using the albedo previously calculated. We evaluated the root mean square difference (RMS) between the rendered and original images, when the synthesized image is created using the albedo estimated along with the other parameters and when the synthesized image is created using albedo estimated by the other subset. We repeated a similar procedure for DB2, by composing an experimental subset using additional images not used in the estimation. The RMS error is calculated for intensities scaled between zero and one applied over the three color channels. It is worth to mention that, in this computation, we only use the pixels enclosed by the silhouette of the mesh projected on the image plane, and not the complete image. If $\mathbf{B}_k$ represent the colors triplet for the pixels k and $\hat{\mathbf{B}}_k$ is the corresponding estimation, then the error RMS is defined as:

$$B_{\text{RMS}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left|\mathbf{B}_k - \hat{\mathbf{B}}_k\right|} \ , \text{ where} \tag{5.34}$$

$$\mathbf{B}_k = \begin{bmatrix} B_r^k & B_g^k & B_b^k \end{bmatrix}^\mathsf{T} \ , \ \hat{\mathbf{B}}_k = \begin{bmatrix} \hat{B}_r^k & \hat{B}_g^k & \hat{B}_b^k \end{bmatrix}^\mathsf{T} \ ,$$

and $|\cdot|$ is the norm of the vector. Using DB1, the average RMS error for the case where images were synthesized using the parameters estimated from the training subset is around 1.2% while for the case where we use the previously estimated albedo the RMS error is higher, as expected (around 17% of the full pixel intensity). When using DB2, the RMS error increases, since the original images contain sometimes pedestrians or objects not taken into account in the 3D model. The error values for all the images in both databases are shown in figure 5.6. These results are useful in the process of cross validation of our algorithm.

Figures 5.7 and 5.8 show the original images and their corresponding synthesized versions for DB1 and DB2. In this case images were recreated using the albedo computed on the training group and the illumination and CRF estimated for the test group .

The third column of figure 5.7 represents the computed spherical harmonics coefficients projected on a unit-radius sphere viewed from the same point of view as the original images. An arrow indicates the maximum point, the direction where the magnitude of the illumination is strongest. It was mentioned in section 5.2.1 that albedos and illumination coefficients can only be estimated up to a global scale factor. This is the case for all three color channels. Hence, in order to display RGB illumination models and surface colors, we first have to estimate the ratios of these scales, between color channels. These scales are calculated by selecting a portion of the sky and projecting its pixels on the sphere. We use the geometric calibration to find where the manually selected region is projected on the unit sphere. Then, the right scale for each channel is found by fitting the spherical harmonics coefficients to the colors of the projected pixels. Images where the direction of the light source can be deduced form shadows show a correct estimation of the illumination direction. For cloudy

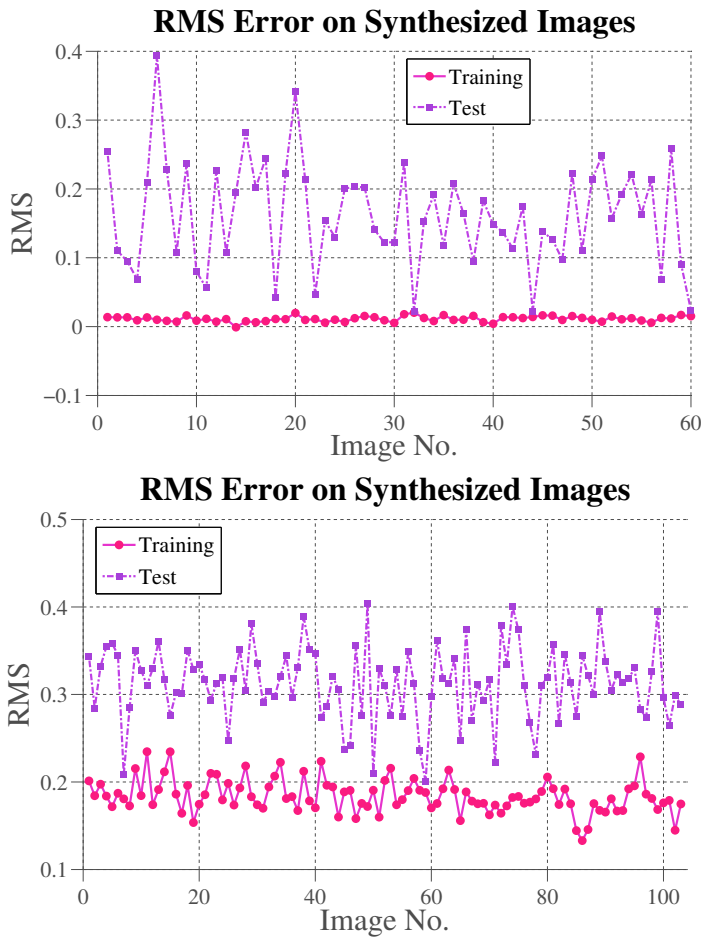


Figure 5.6: RMS error between original and rendered images for database 1 and 2 respectively.

skies, illumination is more uniform (*cf.* third row of figure 5.7) and the maximum is less pronounced. These results suggest that the technique maybe suitable in applications such as image relighting.

5.4 Conclusion and Discussion

This chapter introduced a method to estimate the CRFs and the illumination conditions from an unordered set of images in a joint estimation framework. The computed CRFs show good performance, similar to state–of–the–art methods, with the added value that, in this case, the algorithm also provides information about the illumination of the scene. Results for the CRFs are very close to the estimations found in chapter 4, which is coherent with the formulation presented there. In fact, in both cases we are using the same model for the response functions which limits the space of possible CRF's to the linear combinations formed by a

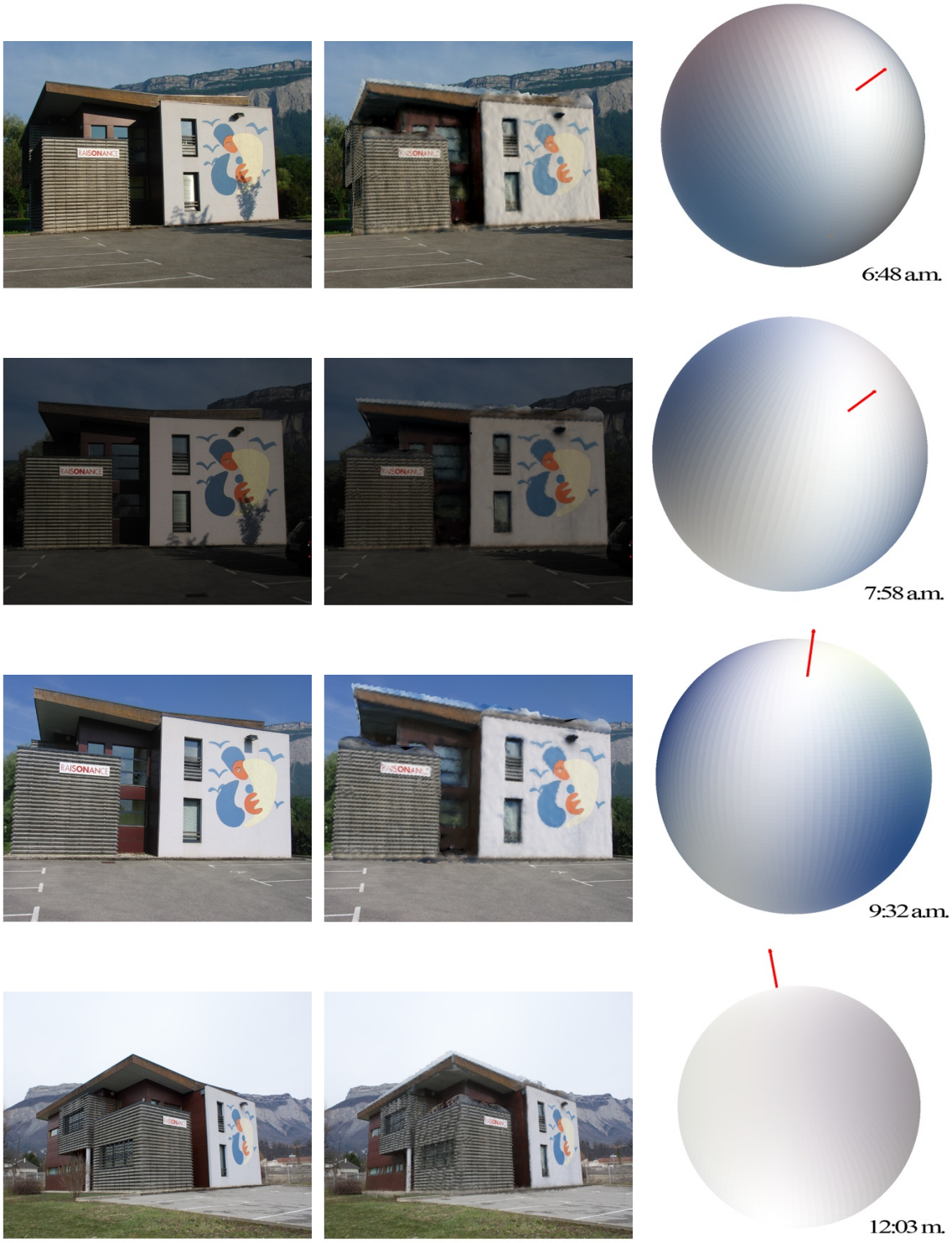


Figure 5.7: The first column shows original images used for reconstruction. Next, we show the 3D model reprojected on the original image with albedos and illumination estimated by our technique. The third column illustrates the illumination estimated for each image. The colored arrow indicates the direction with maximum illumination and the acquisition time is included to give an intuition of the sun's position.

Figure 5.8: The first column shows original images used for reconstruction using an online community photo collection. On the second column we present corresponding rendered images using the estimated CRFs, illumination conditions and albedos. Note that the entire 3D scene model is rendered on top of original images here, without taking into account visibility, which is why for example the wall in the third row's original image is covered by the rendered model.

basis and some coefficients. Also the constraints fixed on the form of the estimated CRFs (monotonically increasing) and the same initial point when starting the non–linear optimizations ensures the convergence of the results in both cases to similar curves.

The use of spherical harmonics for modeling the lighting has allowed us to make a reasonable prediction of the illumination present in the scene during the acquisition time. The representation proposed in [Ramamoorthi 2001b] is a powerful tool to describe the interaction of illumination and surface reflectance. Despite the restrictions on the properties of the surface reflectance (a Lambertian model), this assumption combined with the use of a robust estimation seems to be sufficient to obtain good approximations of the CRF's and the global illumination.

On the other hand, an exact estimation of the albedo might not be possible without an *a priori* knowledge on the gain term given by the camera sensitivity (see the term ρ'_{jk} depending on q_k in equation (5.2), or the more general description of the camera sensitivity model $Q_k(\lambda)$ in equation (2.1)). We are aware that the results on the estimation, specially of the illumination and the albedo, can be improved using a more "personalized" model for the camera sensitivities. The challenge, for the moment, is to formalize a parametric model for the camera which includes this phenomenon (we shall discuss further this point in the next chapter).

Another point of discussion arises around the influence of the 3D model on the calculated values. At the moment, no multi-view reconstruction algorithm can guarantee a 100% accurate reconstruction, and only active reconstruction methods, as structured lighting or laser reconstruction generate, in general, reliable models. However the combination of the two methods used in this work for the 3D reconstruction have shown acceptable results when applying to outdoor scenes. First, the camera calibration obtained using [Snavely 2008] generates camera projection matrices most of the time coherent with the input images. In our experiments, we found an approximative ratio of 1/200 cases of failure for the estimation of the camera projection. Those images were removed. On the side of the dense reconstruction performed by [Furukawa 2010], the method claims to deliver one of the best performances on the evaluation done by the Middlebury database [Seitz 2006]. It has also proven to be a very efficient algorithm for the reconstruction of large scale outdoor scenes. Despite these nice characteristics, there are always artifacts in the recovered 3D models. Most of these critical regions are usually associated to the edges of the structure, where discontinuities are present, or where the information obtained from the images is poor, or even null (the roof of the building in figure 5.4, for example). We justify the good performance of our algorithm, under these conditions, particularly when estimating the parameters related to the whole image (CRF and illumination), by the relatively small number of faulty normals, compared with the large amount of points presenting a satisfactory estimation of the normal. We believe also that our robust estimation helps to discard these outliers by assigning low weights to the wrong points. On the contrary, the albedo estimation for these points is critical, since they depend closely on the normal of the surface element.

There are also some considerations about the size of the non–linear optimization prob-

lem, and the benefits of performing a joint estimation. It is widely accepted that the computation of all the parameters at the same time could bring some advantages, since all the unknown values are changing dynamically and the direction of search uses all the information available in order to calculate the gradient towards a local minimum. However, at the present, if we improve the details on the 3D model or if we use a larger amount of images, the number of parameters to estimate increases and the computational resources of the system might reach their limits. In those cases the estimation can be formulated as a bilinear inference problem by finding a Bayesian probability for the parameters involving the camera and the illumination, and doing similar for the surface reflectance. A similar approach is proposed in [Romeiro 2010] with good results.

Concerning the reflectance properties of the surface, it is also desirable to introduce a more general BRDF model. The number of parameters to compute may increase dramatically because the illumination and the reflectance interact over all the directions of the upper hemisphere centered at the normal of the surface point. We shall extend this discussion during the description of the future works section in the next chapter.

CHAPTER 6

Conclusions and Perspectives

Contents

Conclusion (en francais). La popularisation des appareils photos numériques et la propagation du web 2.0 ont eu un impact major sur la société de la première décennie du siècle XXI. La façon d'observer, capturer et partager le monde à changé grâce aux outils disponibles dans la vie quotidienne. Le stockage et archivage des images provenant de ces appareils photos a donné naissance aux grandes collectionnes photos, c'est a dire, bases de données contenant de milliers de photographies, dans plusieurs cas, d'objets bien connus, registrées depuis différents points de vues avec une palette très hétérogène des conditions. Pendant le déroulement de cette thèse nous avons proposé un groupe de solutions pour profiter de la vaste gamme d'apparences présentées dans ces collections. Toutefois, certaines questions restent particulièrement difficiles.

Déduire information à partir d'images est considéré comme un problème mal posé en raison du nombre élevé d'inconnues et les interactions complexes entre eux. Cette thèse a abordé le problème de deux points de vue : d'abord, en comparant une région commune de toutes les images en plein air : le ciel. Pour cela, des techniques d'analyse d'images sont mises en œuvre afin de comparer les similitudes entre les paires d'images. A fin d'étendre la compréhension de la problématique, nous nous sommes rendu compte que additionalement à l'etude et comparaison de l'apparence, nous avons besoin des autres informations disponibles, y compris la géométrie des objets contenus dans la scène dessinée par les images. Pour cette raison, une seconde formulation a été proposée, en exploitant une autre caractéristique commune de la collection d'images : la structure 3D et les propriétés de réflectance correspondant a cette structure.

Grâce à des méthodes présentées dans ce document, nous avons montré que l'hétérogénéité des collections de photos peut être considérée comme un vrai avantage. Il est possible de bénéficier de toute la richesse de l'apparence que présente les larges collections d'images, si nous formulons des hypothèses raisonnables et si nous utilisons les bons outils pour modéliser tous les phénomènes impliqués.

6.1 Conclusion

With the popularization of digital cameras and the spread use of the web 2.0, the ways of seeing, capturing and sharing the world have had a major impact on the society. Every person with a camera, professional or casual photographer, every Internet surfer, can contribute with his personal point of view of the reality. Putting together all these graphical expressions in the same virtual space creates highly heterogeneous image collections. These datasets, in most of the cases, exhibit a great variety of appearances when imaging the same scene. The increasing amount of available information gives us a plus, over the also increasing noise introduced by them. In this work we propose a group of solutions to take advantage of the wide range of appearance presented in photo collections. However, some particular issues remain challenging.

Deducting information from images is considered an ill-posed problem due to the high number of unknowns and the complex interactions between them. This thesis addressed the problem from two points of view: first by comparing a common region of all the outdoor images: the sky. Image analysis techniques are implemented to compare the similarities between image pairs. But going beyond an appearance comparison requires also the use of all the available information, including the 3D clues that can be extracted from images. For this reason, a second formulation was done, by exploiting another common characteristic of the image collection: the 3D structure and the reflectance properties of the material composing the surface.

Through the methods presented in this document we showed that the heterogeneity of photo collections can be considered as an advantage. It is possible to benefit from all the richness of appearance that large image datasets exhibit, if we state reasonable assumptions and if we count with the right tools to model all the involved phenomena.

6.2 Summary of the proposed methods and contributions

Each of the methods presented to extract photometric information from photo collections has important advantages and disadvantages. During the comparison of images based on the sky appearance, the proposed 3–steps pipelines gather a group of techniques to bring important results. These methods particularly allow us to understand the importance of incorporating the information of multiple images in the problem. Using only planar information, the limits of the algorithms are given by the information contained on the pixels of the images. It seems impossible to infer extra information without having a prior knowledge of the cameras or the acquisition conditions. From a purely technical point of view the technique using a GMM and KL divergence presents various advantages over the classical way of comparing histograms using the Earth Mover's Distance. The capacity of automatically selecting the number of parameters for the model is a plus, and according to our experiments, the results are consistent.

In order to reach the goal of this thesis, the second part of the work consisted in incorporating geometric information to our problem, by calculating a 3D scene structure using the images belonging to the photo collection. A 3D model, the image collection, and the geometric calibration of the cameras are the inputs for our algorithms. We proposed to simulate the image formation process by using models for all the entities involved. These models where represented with different levels of complexity, first by using a simple illumination model and then by applying a general global illumination framework. The exposed algorithms are based on two strong constraints: the first one establishes that the radiometric calibration of the camera can be modeled using the EMOR model. And second, the assumption of a Lambertian representation of the surface reflectance. The main contributions of our work are listed below:

- First, a complete model to describe the sky as a mixture of Gaussian distributions is proposed. These models are suitable for finding similar appearance images in large photo collections using the KL divergence. Obtained results present better performances than those ones exhibited in traditional methods for comparing histograms.

- We estimated the camera response function automatically for all the images belonging to a photo collection. This is a novel approach to estimate the radiometric calibration, since in the past, traditional methods used a calibration target inserted on the scene, multiple exposure images (taken from the same point of view) or specific artifacts in single images. The last mentioned kind of algorithms can not be easily generalized because it requires of a manual tuning parameter to find optimum results.

- The full estimation of illumination, camera response function and surface reflectance is a completely novel approach. Before this work, the image–based estimation of the illumination and the surface reflectance had usually assumed a linear (and/or known) camera response function, which for the case of photo collections compiled from different instances of cameras is not the best choice. Given the high interaction between all the components to calculate, a joint estimation is the most natural way to achieve this goal.

6.3 Perspectives and possible applications

As usual during a research work, besides the solved problems, new questions arise and their responses could be a complement to the methods here presented or the starting point for new research problems, as well.

On the side of image comparison from sky appearance it might be also desirable to integrate information given by other regions of the image, for example shading on planar surfaces, or clues about the illumination raised from the shadows. In [Lalonde 2009], for example, authors use a data–driven approach to estimate a prior and calculate the position of the sun using also information of the shadows on the ground and the shading on vertical

surfaces. After a long training phase (in order to calculate the prior), the results obtained by this algorithm are remarkable. One could imagine to use this kind of clues combined with sky appearance to formulate a more robust comparison algorithm.

A straightforward application can be derived from the image comparison method presented in chapter 3. The photo montage process is a long labour that requires a considerable amount of human efforts, for such a simple task as finding illumination correspondences between to images. For example, given a query image, where we want to discard the background and use only a foreground object, our algorithm is able to propose a series of possible images where it can be inserted, with coherent illumination. Another possible application could be related to the scene based content search. A search algorithm could assign priorities to the images where illumination conditions are similar to those presented on the query image.

Concerning the model for the camera response function used in the formulations, we consider that there exist several ways to improve it, and consequently, ameliorate the reliability of the results. The empirical model introduced by [Grossberg 2004] dates from a time when digital cameras were not majority in the market. The authors of the work used 201 real world CRFs to deduce the whole space of possible functions using a high percentage of analogue cameras, and including only a few digital cameras. Moreover, the behavior of these devices to the actual cameras have radically changed. The problem of modeling the Camera Response Functions of actual cameras is really challenging because of all the internal processes carried out by the embedded electronics. First of all, the scope of what a response function must represent. Does the Camera Response Function only refer to the relation of the image irradiance and the "raw" output of the sensor? Does it also include all the post-processing changes internally done by the camera? If this is the case, how the white balance, the contrast improvement, the image compression and other processes alter the CRF? What should be the relation of the CRFs, the light spectral distribution and the color channels? Is the CRF dependent on the scene? Is the CRF dependent on the sensor spectral sensitivity? Some of these interrogatives have been partially answered by [Chakrabarti 2009] where the authors propose a camera response model relating the color channels of the image. However, the proposed approach brings us back to the first models proposed for analogue cameras, where the non–linear curve was represented by a polynomial of a high degree. The challenge remains open, and the formulation of a standard model for the current consumer cameras is certainly eagerly awaited by the community.

The algorithms proposed in the second part of this thesis (chapters 4 and 5) can also be included in the framework of a full 3D reconstruction process. In addition to the estimation of the CRFs, the illumination and the albedos, one can consider the normals of the 3D surface element as approximations to the real values. In this case, the previous step of geometric reconstruction generates the initial values. One can formulate an optimization strategy to tune up at the same time the normals and improve the quality of the geometric reconstruction. Additionally, the inclusion of advanced models for the BDRF might also be useful in some scenarios. One alternative consists on using a catalogue of materials to represent common

BRDFs of particular objects (the components of a building, for example). Another option is to insert an advanced model for the material reflectance by representing the BRDF as a linear combination of basis functions. They can be learned through non-negative matrix factorization (NMF) of a database containing typical samples of the reflectance properties for different materials. This strategy is proposed by [Romeiro 2010] and could also be valid in our case. The inlcusion of more detailed models for the illumination may be also envisaged. For this, one alternative can be to fuse the estimated illumination with relevant information deduced from the image matadata (wheather forecast, time of day, season, geolocalisation, etc.). Finally, given that the 3D structure is readly available, one can consider to find the visibility on the light path and include this information for the estimation. All these alternatives could improve the robsutness of the proposed approaches and enlarge also the scope of this work.

APPENDIX A

Extended results for Chapter 3

A subjective evaluation could be done by observing the images closest to a particular image. Algorithms described in chapter 3 are used to compared one particular image to 1000 images in the database according to the metric described: $\mathsf{KL}_{\text{simple}}$, KL_{app} and EMD. The first column corresponds to the query image and the following columns present the images with lowest values, increasing from left to right. In this figure, the similarity between the sky of the query image and the skies of the found images is evident.

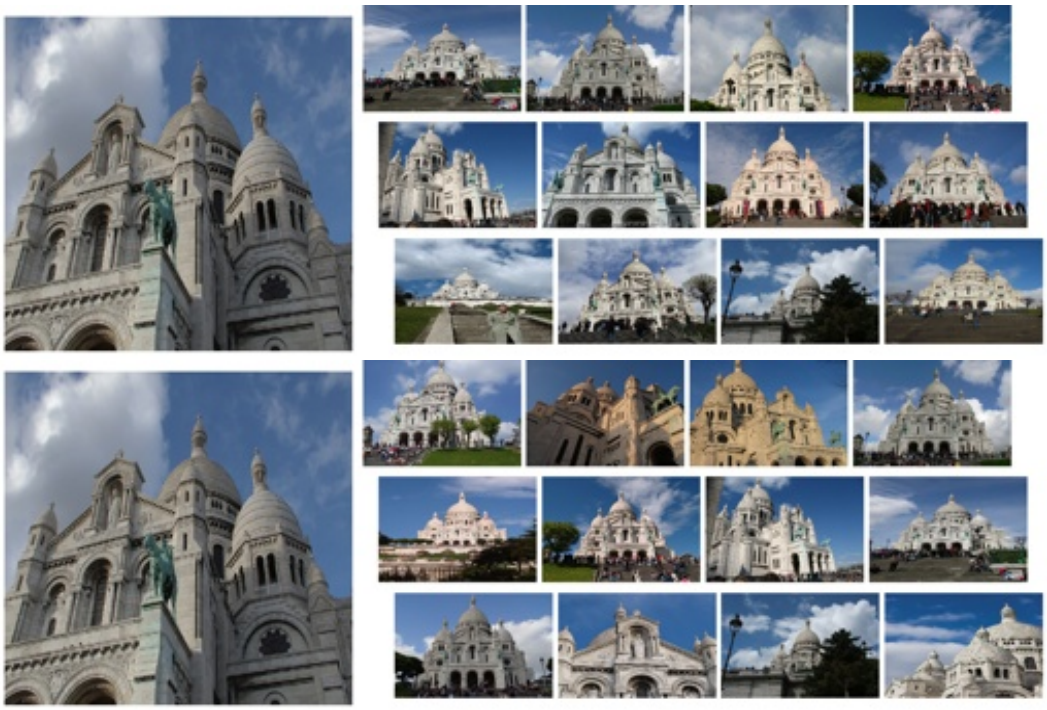


Figure A.1: Results database Sacrecoeur

Figure A.2: Results database Sacrecoeur

Figure A.3: Results database Sacrecoeur

Figure A.4: Results database Pisa

Figure A.5: Results database Pisa

Figure A.6: Results database Pisa

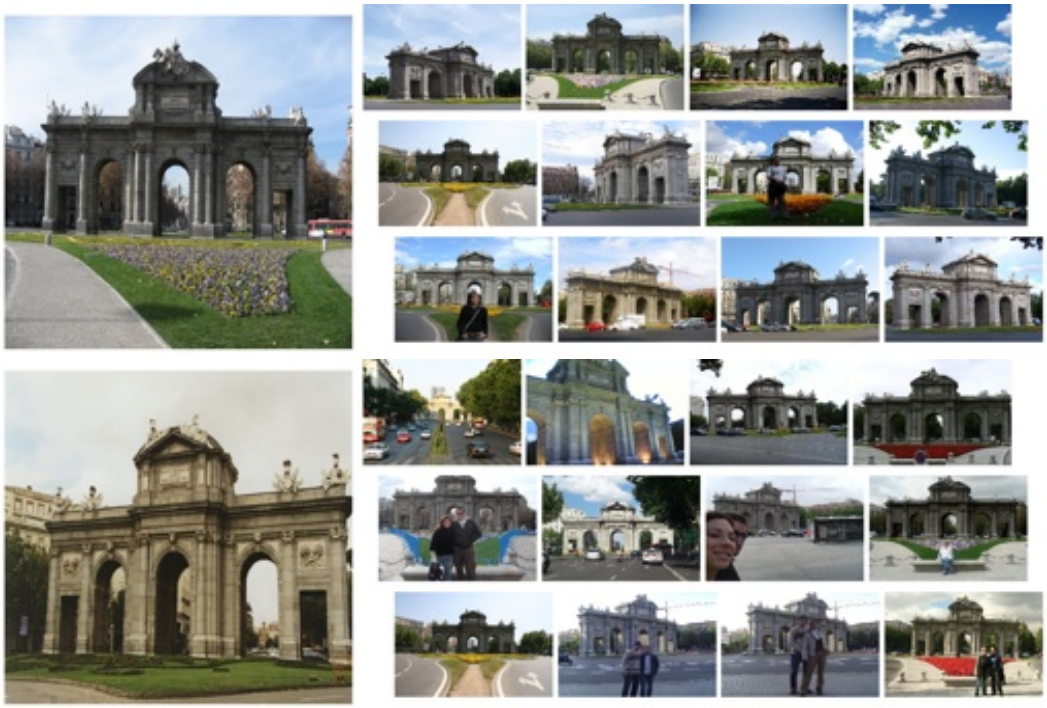


Figure A.7: Results database Palcala

Figure A.8: Results database Palcala

Figure A.9: Results database Palcala

Bibliography

[Adelson 1991] Edward Adelson. *The plenoptic function and the elements of early vision*. Computational Vodels of Visual Processing, Jan 1991. (Cited on page 16.)

[Akaike 1974] Hirotugu Akaike. *A new look at the statistical model identification*. Automatic Control, IEEE Transactions on, vol. 19, no. 6, pages 716 – 723, dec 1974. (Cited on page 37.)

[Basri 2003] Ronen Basri and David Jacobs. *Lambertian reflectance and linear subspaces*. IEEE Transactions on Pattern Analysis and Machine Learning, Jan 2003. (Cited on pages 20, 71 and 77.)

[Bay 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars and Luc Van Gool. *Speeded-Up Robust Features (SURF)*. Computer Vision and Image Understanding, vol. 110, no. 3, pages 346 – 359, 2008. Similarity Matching in Computer Vision and Multimedia. (Cited on page 14.)

[Bilmes 1998] Jeff Bilmes. *A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models*. International Computer Science Institute, Jan 1998. (Cited on page 27.)

[Bishop 2006] Christopher M. Bishop. Pattern recognition and machine learning. Springer Science, Business Media, LLC, 2006. (Cited on pages 27 and 37.)

[Blinn 1976] James Blinn and Martin Newell. *Texture and reflection in computer generated images*. Communications of the ACM, Jan 1976. (Cited on page 19.)

[Bouguet 1998] Jean-Yves Bouguet and Pietro Perona. *Camera calibration from points and lines in dual-space geometry*. Citeseer, Jan 1998. (Cited on page 14.)

[Cabral 1987] B Cabral and N Max.... *Bidirectional reflection functions from surface bump maps*. ACM SIGGRAPH Computer ..., Jan 1987. (Cited on page 77.)

[Chakrabarti 2009] Ayan Chakrabarti, Daniel Scharstein and Todd Zickler. *An Empirical Camera Model for Internet Color Vision*. Proceedings of the British Machine Vision Conference, pages xx–yy, 2009. (Cited on pages 22 and 106.)

[Chang 1996] Young-Chang Chang and J Reid. *RGB calibration for color image analysis in machine vision*. Image Processing, IEEE Transactions on, vol. 5, no. 10, pages 1414 – 1422, Oct 1996. (Cited on page 21.)

[Cook 1981] Robert Cook and Kenneth Torrance. *A reflectance model for computer graphics*. ACM SIGGRAPH Computer Graphics, Jan 1981. (Cited on page 18.)

[Criminisi 2000] Antonio Criminisi, Ian Reid and Andrew Zisserman. *Single view metrology*. International Journal of Computer Vision, Jan 2000. (Cited on page 14.)

[C.T 1994] Committee C.T. *Spatial distribution of daylight - luminance distributions of various reference skies.* Rapport technique CIE-110-1994, Commission Internationale de l'Éclairage (CIE), 1994. (Cited on page 34.)

[C.T 2002] Committee C.T. *Colour Appearance Model for Colour Management Applications*. Rapport technique CIE-TC8-01, Commission Internationale de l'Éclairage (CIE), 2002. (Cited on page 34.)

[Debevec 1996] Paul Debevec, Camillo Taylor and Jitendra Malik. *Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach*. SIGGRAPH Proceeding of the conference on Computer Graphics and Interactive Techniques, Jan 1996. (Cited on pages 16 and 19.)

[Debevec 1997] Paul Debevec and Jitendra Malik. *Recovering high dynamic range radiance maps from photographs*. SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, Aug 1997. (Cited on pages 21, 67, 69 and 70.)

[Debevec 2004] Paul Debevec, Chris Tchou, Andrew Gardner, Tim Hawkins, Charis Poullis, Jessi Stumpfel, Andrew Jones, Nathaniel Yun, Per Einarsson and Therese Lundgren. *Estimating surface reflectance properties of a complex scene under captured natural illumination*. Conditionally Accepted to ACM Transactions on Graphics, 2004. (Cited on page 19.)

[Delaunoy 2008] Amaël Delaunoy, Emmanuel Prados, Pau Gargallo, Jean Philippe Pons and Peter Sturm. *Minimizing the Multi-view Stereo Reprojection Error for Triangular Surface Meshes*. certis.enpc.fr, Jan 2008. (Cited on page 15.)

[Dempster 1977] A Dempster and N Laird. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society., Jan 1977. (Cited on page 27.)

[Dorsey 2007] Julie Dorsey, Holly Rushmeier and François Sillion. Digital modeling of material appearance. Elsevier, 2007. (Cited on page 16.)

[Esteban 2008] Carlos Hernández Esteban, George Vogiatzis and Roberto Cipolla. *Multiview Photometric Stereo*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, no. 3, pages 548–554, 2008. (Cited on page 15.)

[Faugeras 1993] Olivier Faugeras. Three-dimensional computer vision: a geometric viewpoint. MIT Press, Jan 1993. (Cited on page 13.)

[Finlayson 2006] G Finlayson, S Hordley, Cheng Lu and M Drew. *On the removal of shadows from images*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 1, pages 59 – 68, 2006. (Cited on pages 17 and 72.)

[Fischler 1981] Martin A Fischler and Robert C Bolles. *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*. Commun. ACM, vol. 24, pages 381–395, Jun 1981. (Cited on page 14.)

[Freund 1995] Yoav Freund and Robert Schapire. *A desicion-theoretic generalization of online learning and an application to boosting*. Computational learning theory, Jan 1995. (Cited on page 34.)

[Furukawa 2009] Yasutaka Furukawa and Jean Ponce. *Accurate, Dense, and Robust Multi-View Stereopsis*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009. (Cited on page 15.)

[Furukawa 2010] Yasutaka Furukawa, Brian Curless and Steven M Seitz.... *Towards internet-scale multi-view stereo*. IEEE Conference on Computer Vision and Pattern Recognition, Jan 2010. (Cited on pages 15, 66 and 101.)

[Gargallo 2008] Pau Gargallo. *Contributions to the Bayesian Approach to Multi-View Stereo*. PhD thesis, Institut National Polytechnique de Grenoble (INPG), Feb 2008. (Cited on page 15.)

[Gevers 2001] Theo Gevers. *Color in Image Search Engines*. In Principles of Visual Information Retrieval. Springer–Verlag, London, 2001. (Cited on page 35.)

[González 2008] Rafael C. González and Richard Eugene Woods. Digital image processing. Pearson, Prentice Hall, 2008. (Cited on page 16.)

[Groemer 1996] H. Groemer. Geometric applications of fourier series and spherical harmonics. New York: Cambridge University Press, 1996. (Cited on pages 77 and 79.)

[Grossberg 2002] Michael D Grossberg and Shree K Nayar. *What can be known about the radiometric response from images?* LECTURE NOTES IN COMPUTER SCIENCE, Jan 2002. (Cited on page 24.)

[Grossberg 2003] Michael D Grossberg and Shree K Nayar. *What is the Space of Camera Response Functions?* Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference onSun, vol. 2, page 602, 2003. (Cited on pages 22, 56, 68, 69 and 70.)

[Grossberg 2004] Michael D Grossberg and Shree K Nayar. *Modeling the space of camera response functions*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 26, no. 10, pages 1272 – 1282, Oct 2004. (Cited on pages 22, 23, 51, 67, 84 and 106.)

[Harris 1988] C Harris and M Stephens. *A Combined Corner and Edge Detection*. Proceedings of The Fourth Alvey Vision Conference, pages 147–151, 1988. (Cited on page 14.)

[Hartley 2003] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge University Press, ISBN: 0521540518, second édition, 2003. (Cited on pages 13, 14 and 90.)

[Hays 2007] James Hays and Alexei Efros. *Scene Completion Using Millions of Photographs*. ACM Transactions on Graphics (SIGGRAPH 2007), vol. 26, no. 3, 2007. (Cited on page 41.)

[Heckbert 1992] P Heckbert. *Introduction to global illumination*. Global Illumination Course SIGGRAPH, Jan 1992. (Cited on page 19.)

[Heidrich 1998] Wolfgang Heidrich and Hans-Peter Seidel. *View-independent environment maps*. Proceedings of the ACM SIGGRAPH, Jan 1998. (Cited on page 20.)

[Hershey 2007] John Hershey and Peder Olsen. *Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models*. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol. 4, pages IV–317 – IV–320, Mar 2007. (Cited on page 40.)

[Hoiem 2005] Derek Hoiem, Alexei Efros and Martial Hebert. *Geometric context from a single image*. Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 1, pages 654 – 661 Vol. 1, Sep 2005. (Cited on page 34.)

[Horn 1986] Berthold Horn. Robot vision. The MIT Press, McGraw–Hill Book Company, 1986. (Cited on page 12.)

[Igawa 2004] Norio Igawa, Y Koga, T Matsuzawa and H Nakamura. *Models of sky radiance distribution and sky luminance distribution*. Solar Energy, Jan 2004. (Cited on page 24.)

[Ilie 2005] Adrian Ilie and Greg Welch. *Ensuring color consistency across multiple cameras*. IEEE International Conference on Computer Vision (ICCV), Jan 2005. (Cited on page 21.)

[Ishimaru 1991] A Ishimaru. *Electromagnetic wave propagation, radiation, and scattering*. Prentice Hall, Jan 1991. (Cited on page 16.)

[Judd 1964] Deane Judd, David MacAdam, Gutnter Wyszecki, H Budde and H. R. Condit. *Spectral distribution of typical daylight as a function of correlated color temperature*. JOSA, Jan 1964. (Cited on page 25.)

[Kazhdan 2006] Michael Kazhdan, Matthew Bolitho and Hugues Hoppe. *Poisson surface reconstruction*. In Proceedings of the fourth Eurographics symposium on Geometry processing, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association. (Cited on page 66.)

[Kim 2008] Seon Joo Kim, J Frahm and Marc Pollefeys. *Radiometric calibration with illumination change for outdoor scene analysis*. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1 – 8, 2008. (Cited on page 71.)

[Kuthirummal 2008] Sujit Kuthirummal, Aseem Agarwala, Dab Goldman and Shree K Nayar. *Priors for Large Photo Collections and What They Reveal about Cameras*. Proceedings of the European Conference in Computer Vision (ECCV), pages 74–87, Oct 2008. (Cited on page 21.)

[Lafortune 1997] Eric Lafortune, Sing-Choong Foo, Kenneth Torrance and Donald Greenberg. *Non-linear approximation of reflectance functions*. Proceedings of the SIGGRAPH conference, Jan 1997. (Cited on page 18.)

[Lalonde 2007] Jean-François Lalonde, Derek Hoiem, Alexei Efros, Carsten Rother and John Winn. *Photo clip art*. ACM transactions on Graphics (SIGGRAPH), Jan 2007. (Cited on page 35.)

[Lalonde 2008] Jean-François Lalonde, Srinivasa Narasimhan and Alexei Efros.... *What does the sky tell us about the camera*. European Conference on Computer Vision, 2008. (Cited on pages 24 and 35.)

[Lalonde 2009] Jean-François Lalonde, Alexei Efros and Srinivasa Narasimhan. *Estimating Natural Illumination from a Single Outdoor Image*. ICCV 2009, 2009. (Cited on page 105.)

[Lalonde 2011] Jean-François Lalonde. *Understanding and Recreating Visual Appearance Under Natural Illumination*. PhD thesis, Carnegie Mellon University, Jan 2011. (Cited on pages 24 and 32.)

[Lambert 1760] Johann Heinrich Lambert. Photometrie: Photometria, sive de mensura et gradibus luminis, colorum et umbrae. the Bavarian State Library, 1760. (Cited on pages 1 and 3.)

[Levenberg 1944] K Levenberg. *A method for the solution of certain problems in least squares*. Quarterly of Applied Mathematics, Jan 1944. (Cited on page 25.)

[Liebowitz 1998] D Liebowitz and Andrew Zisserman. *Metric rectification for perspective images of planes*. Computer Vision and Pattern Recognition (CVPR), Jan 1998. (Cited on page 14.)

[Lin 2004] Steve Lin, Jinwei Gu, S Yamazaki and Heung-Yeung Shum;. *Radiometric calibration from a single image*. Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2, pages II–938 – II–945 Vol.2, 2004. (Cited on pages 21, 24, 69, 70, 71 and 73.)

[Lin 2005] Steve Lin and L Zhang. *Determining the radiometric response function from a single grayscale image*. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, pages 66 – 73 vol. 2, 2005. (Cited on pages 21 and 71.)

[Ling 2007] Haibing Ling and Kanzunori Okada. *An efficient earth mover's distance algorithm for robust histogram comparison*. IEEE transactions on pattern analysis and recognition, Jan 2007. (Cited on pages 39 and 40.)

[Lourakis 2009] M.I.A Lourakis and A.A Argyros. *SBA: A software package for generic sparse bundle adjustment*. ACM Transactions on Mathematical Software (TOMS), vol. 36, no. 1, pages 1–30, 2009. (Cited on pages 14, 27, 86, 90 and 92.)

[Love 1997] Robert Love. *Surface reflection model estimation from naturally illuminated image sequences*. PhD thesis, The University of Leeds, School of Computer Studies, 1997. (Cited on page 24.)

[Lowe 2004] David G Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. Int. J. Comput. Vision, vol. 60, no. 2, pages 91–110, 2004. (Cited on page 14.)

[Luong 2002] Q Luong, P Fua and Y Leclerc. *The radiometry of multiple images*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 1, pages 19 – 33, Jan 2002. (Cited on pages 19, 86 and 87.)

[MacRobert 1967] T. M. MacRobert and I. N. Sneddon. Spherical harmonics: An elementary treatise on harmonic functions, with applications. Oxford, England: Pergamon Press, 1967. (Cited on page 77.)

[Mann 1995] S Mann and R Picard. *On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures*. Proceedings of IS&T, Soc. for Imaging Science and Technology 46th Ann. Conf., pages 442—448, Jan 1995. (Cited on page 22.)

[Marquardt 1963] D Marquardt. *An algorithm for least-squares estimation of nonlinear parameters*. Journal of the society for Industrial and Applied ..., Jan 1963. (Cited on pages 14 and 26.)

[Matsushita 2007] Y Matsushita and Steve Lin. *Radiometric Calibration from Noise Distributions*. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1 – 8, 2007. (Cited on page 71.)

[McLachlan 1997] Geoffrey McLachlan. The em algorithm and extensions. John Wiley and Sons INC, Wiley Interscience, 1997. (Cited on page 27.)

[Miller 1984] Gene S Miller and C Robert Hoffman. *Illumination and Reflection Maps: Simulated Objects In Simulated and Real Environments*, 1984. (Cited on page 19.)

[Mitsunaga 1999] Toomo Mitsunaga and Shree K Nayar. *Radiometric self calibration*. Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 1, Jul 1999. (Cited on pages 21, 22 and 68.)

[Moravec 1980] Hans Moravec. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. Rapport technique CMU-RI-TR-80-03, Carnegie Mellon University Robotics Institue, Sep 1980. (Cited on page 14.)

[Moré 1978] Jorge J. Moré. *The Levenberg-Marquardt algorithm: implementation and theory*. In G Watson, editeur, Numerical analysis - Lecture Notes in Mathematics. Springer Berlin / Heidelberg, Jan 1978. (Cited on page 26.)

[Muja 2009] Marius Muja and David G. Lowe. *Fast approximate nearest neighbors with automatic algorithm configuration*. International Conference on Computer Vision Theory, Jan 2009. (Cited on page 14.)

[Muselet 2004] Damien Muselet, Ludovic Macaire, Pierre Bonnet and Jack-Gérard Postaire. *Reconnaissance d'objets sous éclairage non contrôlé par l'intersection entre histogrammes couleur spécifiques*. Traitement du Signal, Numéro spécial sur L'image numérique couleur, Jan 2004. (Cited on page 29.)

[Muselet 2005] Damien Muselet. *Reconnaissance automatique d'objets sous éclairage non contrôlé par analyse d'images couleur*. PhD thesis, Université de Sciences et Technologies de Lille 1, July 2005. (Cited on page 30.)

[Ng 2004] R Ng, Ravi Ramamoorthi and Pat Hanrahan. *Triple product wavelet integrals for all-frequency relighting*. ACM SIGGRAPH 2004 Papers, pages 477–487, 2004. (Cited on page 77.)

[Ng 2007] Tian-Tsong Ng, Shih-Fu Chang and Mao-Pei Tsui;. *Using Geometry Invariants for Camera Response Function Estimation*. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1 – 8, 2007. (Cited on page 21.)

[Nicodemus 1992] F E Nicodemus, J C Richmond, J J Hsia, I W Ginsberg and T Limperis. *Geometrical considerations and nomenclature for reflectance*. In Lawrence B. Wolff, Steven A. Shafer and Glenn Healey, editeurs, Radiometry, pages 94–145. Jones and Bartlett Publishers, Inc., 1992. (Cited on page 17.)

[Nister 2006] David Nister and Henrik Stewenius. *Scalable recognition with a vocabulary tree*. Computer Vision and Pattern Recognition, Jan 2006. (Cited on page 14.)

[Pascale 2006] Danny Pascale. *RGB Coordinates of the Macbeth ColorChecker*. Rapport technique, The BabelColor Company, Jun 2006. (Cited on page 64.)

[Pele 2009] Ofir Pele. *Fast and robust earth mover's distances*. Computer Vision, Jan 2009. (Cited on page 30.)

[Perez 1993] R Perez and R Seals. *All-weather model for sky luminance distribution–preliminary configuration and validation*. Solar Energy, Jan 1993. (Cited on page 24.)

[Phong 1975] B Phong. *Illumination for computer generated pictures*. Communications of the ACM, Jan 1975. (Cited on page 18.)

[Pratt 2001] W Pratt. *Digital image processing: PIKS inside*. John Wiley & Sons Inc, Jan 2001. (Cited on page 16.)

[Ramamoorthi 2001a] Ravi Ramamoorthi. *Modeling illumination variation with spherical harmonics*. Face Processing: Advanced Modeling and Methods, 2001. (Cited on pages 71 and 77.)

[Ramamoorthi 2001b] Ravi Ramamoorthi and Pat Hanrahan. *A signal-processing framework for inverse rendering*. SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, Aug 2001. (Cited on pages 20, 77, 78, 82 and 101.)

[Ramamoorthi 2002] Ravi Ramamoorthi. *Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 10, pages 1322 – 1333, Oct 2002. (Cited on pages 71 and 77.)

[Ren 2003] Xiaofeng Ren and Jitendra Malik. *Learning a classification model for segmentation*. International Conference on Computer Vision, Jan 2003. (Cited on page 34.)

[Romeiro 2010] Fabiano Romeiro and Todd Zickler. *Blind reflectometry*. European Conference on Computer Vision (ECCV), Jan 2010. (Cited on pages 25, 71, 102 and 107.)

[Rubner 1998] Yossi Rubner, Carlo Tomasi and Leonidas Guibas. *A metric for distributions with applications to image databases*. Computer Vision, 1998. Sixth International Conference on, pages 59 – 66, 1998. (Cited on pages 39 and 40.)

[Rubner 2000] Yossi Rubner and Carlo Tomasi. *The earth mover's distance as a metric for image retrieval*. International Journal of Computer Vision, Jan 2000. (Cited on page 29.)

[Schlick 1994] Christophe Schlick. *A Survey of Shading and Reflectance Models*. Computer Graphics forum, vol. 13, pages 121–131, 1994. (Cited on page 19.)

[Seitz 2006] Steven M Seitz, Brian Curless, James Diebel and D Scharstein. *A comparison and evaluation of multi-view stereo reconstruction algorithms*. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Jan 2006. (Cited on pages 15 and 101.)

[Shi 1994] Jianbo Shi and Carlo Tomasi. *Good features to track*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 593 –600, Jun 1994. (Cited on page 14.)

[Sillion 1994] François Sillion and Claude Puech. Radiosity and global illumination. Elsevier, 1994. ISBN 1-558. (Cited on page 16.)

[Slater 1998] D Slater. *What is the spectral dimensionality of illumination functions in outdoor scenes?* Computer Vision and Pattern Recognition (CVPR), Jan 1998. (Cited on page 24.)

[Snavely 2008] Noah Snavely, Steven M Seitz and Richard Szeliski. *Modeling the world from internet photo collections*. International Journal of Computer Vision, Jan 2008. (Cited on pages 12, 14, 15, 66, 88 and 101.)

[Stein 1971] Elias M. Stein and Guido Weiss. Introduction to fourier analysis on euclidean spaces. Princeton University Press, 1971. (Cited on page 20.)

[Sturm 1999] Peter Sturm. *On plane-based camera calibration: A general algorithm, singularities, applications*. Computer Vision and Pattern Recognition, Jan 1999. (Cited on page 14.)

[Sunkavalli 2008] Kalyan Sunkavalli, Fabiano Romeiro, Wojciech Matusik, Todd Zickler and Hanspeter Pfister. *What do color changes reveal about an outdoor scene?* Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference onSun, pages 1–8, Jun 2008. (Cited on page 25.)

[Szeliski 2009] R Szeliski. *Computer Vision: Algorithms and Applications*. research.microsoft.com, Jan 2009. (Cited on page 13.)

[Takamatsu 2008] J Takamatsu, Y Matsushita and K Ikeuchi. *Estimating camera response functions using probabilistic intensity similarity*. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1 – 8, 2008. (Cited on pages 21 and 71.)

[Triggs 2000] Bill Triggs, Philip McLauchlan, Richard Hartley and Andrew Fitzgibbon. *Bundle adjustment—a modern synthesis*. Vision algorithms: Theory and Practice, Jan 2000. (Cited on page 14.)

[von Goethe 1840] Johann Wolfgang von Goethe. Theory of colours. MIT Press, Jan 1840. (Cited on pages 2 and 3.)

[Ward 1992] G Ward. *Measuring and modeling anisotropic reflection*. ACM SIGGRAPH Computer Graphics, Jan 1992. (Cited on page 18.)

[Ward 1994] G Ward. *The RADIANCE lighting simulation and rendering system*. Proceedings of the 21st annual conference on . . . , Jan 1994. (Cited on page 24.)

[Weiss 2009] Y Weiss and A Torralba. *Spectral hashing*. Advances in neural information, Jan 2009. (Cited on page 14.)

[Wilburn 2008] B Wilburn, Hui Xu and Y Matsushita. *Radiometric calibration using temporal irradiance mixtures*. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1 – 7, 2008. (Cited on page 71.)

[Yuan 2000] Y Yuan. *A review of trust region algorithms for optimization*. ICIAM 99: proceedings of the Fourth International, Jan 2000. (Cited on page 26.)

[Zhang 1997] Z Zhang. *Parameter estimation techniques: A tutorial with application to conic fitting*. Image and vision Computing, Jan 1997. (Cited on pages 87 and 89.)

[Zickler 2009] Todd Zickler. *Principles of Appearance Acquisition and Representation*. Foundations and Trends in Computer Graphics and Vision, pages 1–119, Aug 2009. (Cited on page 16.)

[Zmura 1991] Michel DŹmura. *Shading ambiguity: reflectance and illumination*. In Computational Models of Visual Processing, pages 187–207. Cambridge: MIT, 1991. (Cited on page 77.)