Università degli Studi di Pisa

Dipartimento di Informatica
Dottorato di Ricerca in Informatica

Ph.D. Thesis

Surface Appearance Estimation from Video Sequences

Gianpaolo Palma

Supervisor
Dr. Roberto Scopigno

Dr. Marco Callieri

Referee
Prof. Holly Rushmeier

Prof. Jean-Michel Dischler

Chair
Prof. Pierpaolo Degano

INF/01

May, 2013

To my family,
who has always supported and encouraged my studies.

Abstract

The realistic virtual reproduction of real world objects using Computer Graphics techniques requires the accurate acquisition and reconstruction of both 3D geometry and surface appearance. The ability to play interactively with the reflectance, changing the view and the light(s) direction, is mandatory in most applications. In many cases, image synthesis should be based on real, sampled data: synthetic images should comply with sampled images of the real artwork. Unfortunately, in several application contexts, such as Cultural Heritage (CH), the reflectance acquisition can be very challenging due to the type of object to acquire and the digitization conditions. Although several methods have been proposed for the acquisition of object reflectance, some intrinsic limitations still make reflectance acquisition a complex task for CH artworks: the use of specialized instruments (dome, special setup for camera and light source, etc.) that require to move the artwork from its usual location; the need of highly controlled acquisition environments, such as a dark room, which are difficult to be reproduced in standard environments (such as museums, historical buildings, outdoor locations, etc.); the difficulty to extend to objects of arbitrary shape and size; the high level of expertise required to assess the quality of the acquired surface appearance.

This thesis proposes novel solutions for the acquisition and the estimation of the surface appearance in fixed and uncontrolled lighting conditions with several degree of approximations (from a perceived near diffuse color to a SVBRDF), taking advantage of the main features that differentiate a video sequences from an unordered photos collections: the temporal coherence; the data redundancy; the easy of the acquisition, which allows acquisition of many views of the object in a short time.

Finally, Reflectance Transformation Imaging (RTI) is an example of widely used technology for the acquisition of the surface appearance in the CH field, even if limited to single view Reflectance Fields of nearly flat objects. In this context, the thesis addresses also two important issues in RTI usage: how to provide better and more flexible virtual inspection capabilities with a set of operators that improve the perception of details, features and overall shape of the artwork; how to increase the possibility to disseminate this data and to support remote visual inspection of both scholar and ordinary public.

Acknowledgments

I would like to express my deep gratitude to the supervisors, Dr. Roberto Scopigno and Dr. Marco Callieri, for their precious encouragement and their helpful critics during my Ph.D. I would also like to thank all the people in Visual Computing Lab, who have worked with me over the past few years, giving me assistance and important advices.

Thanks to Prof. Michael Goesele for the opportunity to visit his lab and for the valuable suggestions he gave me to improve my work.

A thanks goes to the referees Prof. Holly Rushmeier and Prof. Jean-Michel Dischler, for their helpful comments and suggestions to improve this thesis, and to Prof. Giuseppe Attardi and Prof. Paolo Ferragina for their work in my PhD committee.

The final thanks, the most important, goes to all the people who have endured me over the years, especially my parents and my sisters who have supported and helped me unconditionally.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The realistic and accurate virtual reproduction of real world objects using Computer Graphics techniques is very important in several application fields. The fundamental steps of this process are the acquisition of the 3D geometry and the reconstruction of the surface appearance of the object. Especially in the Cultural Heritage domain, the acquisition of the appearance is very important for the study and the visualization of the artwork, where the characteristic of the material, the reflectance behavior and the texture offer major perceptual and cognitive hints to the user. In many cases, the ability to play interactively with the reflectance, changing the view and light direction, is more useful than the manipulation of an accurately sampled 3D shape, which is hardly able to capture all the interesting aspects of the artwork. On the other hand, in the CH context, the acquisition can be very challenging due to the type of object to acquire and the digitization conditions. This chapter presents the motivations that lead us to develop novel techniques for the approximation of the surface appearance and outlines the contribution of this thesis to the state of the art.

1.1 Motivation

The methodologies and technologies for the acquisition of accurate digital 3D representations of real objects improved in a considerable manner in the last few years. High quality digital 3D models of real artifacts are needed in many different application fields, such as Industrial Design, Surgery and Computer-Aided Health Care, Cultural Heritage, Architecture, etc. Among these fields, the acquisition of Cultural Heritage (CH) objects is a challenging activity, since it requires acquiring very dense and accurate models of both the shape and the surface appearance of the artwork of interest. The final high-resolution model could be used both for the virtual study and inspection of the artwork and for the presentation to the public. Moreover, CH applications require technologies able to acquire objects of very different scales, from a small jewel to an entire building, with different types of materials (multi material object are very common), and usually in uncontrolled lighting environment. The

usual digitization procedure takes place on-the-field, in environments that cannot be modified and from which the object cannot be moved, for example a museum hall or outdoor.

Concern the methodologies for the acquisition of the shape, consolidated scanning technologies developed for 3D model acquisition allow the reconstruction of very accurate geometries with affordable times and high accuracy. Moreover, the availability of several 3D scanning techniques, each one with different characteristics, let the user choose the better solution in function of the proprieties of the object and of the digitization conditions. These techniques allow acquisitions with different degrees of accuracy, of objects of different size, in almost all acquisition environments (indoor and outdoor) with several lighting conditions.

On the other hand, the acquisition of the object's surface appearance is more complex and the technology is in a more primitive status. The surface appearance depends on the scattering function, which encodes how the surface reflects and transmits the light radiation. This function is completely described by 12 parameters that depend by the light and the view direction and the properties of the material. Since a mathematical formulation of this function does not exist and the measurement of a 12-dimensional function is currently not practical, several attempts to estimate a simplified version of this function have been proposed in the last decade. These simplified functions are obtained by introducing some assumptions on the scattering process that constrain the parameters and the type of reflectance effects that the final estimation is able to reproduce.

Nevertheless, being able to encode a reflection function is not enough. While, for pristine materials, there exist tabulated encoding of the corresponding scattering functions, obtained by using complex sampling devices on small pieces of the target material, for the many materials that exist in the real world it is not easy to find a proper representative of their spatially varying reflectance in an atlas. This is especially true for the many different materials and finishes that characterize a CH artwork. It is very hard to characterize a given surface that could have very different types of patinas and degradation processes that modify its reflectivity. Being able to acquire the surface reflection characteristics of that specific surface is therefore strategical in this context.

Unfortunately, the available methodologies for the acquisition of surface reflection are not yet as mature as the acquisition of the geometry. While precise instruments has been designed to characterize small pieces of uniform materials, the acquisition of complex, spatially varying and highly different types of surfaces is still an unsolved and open problem. The main drawbacks of the state of the art solutions are:

- Lack of flexibility. While several specialistic solutions have been developed for very specific problems (for example the possibility to acquire the human face and to use these measurements to render the face under arbitrary changes in lighting and viewpoint), it is difficult to extend these solutions to more general

problems characterized by different shape, size and surface appearance;

- Acquisition in lab with highly controlled lighting condition, such as dark room. It is very hard to reproduce this lab condition when the lighting environment cannot be modified or the artwork cannot be moved from its environment (e.g. in a museum or outdoor);

- Time required to acquire, calibrate and elaborate huge sets of images. The surface appearance acquisition should not be longer than the shape acquisition and it should require at most some hours even for the more complex artworks;

- High level of expertise to evaluate on site the completeness and quality of the data acquisition (e.g. if the photographic acquisition is dense enough to reconstruct the surface appearance). The development of an automatic methodology to reduce the contribution in time and expertise of the operator would be desirable.

This thesis proposes some innovative solutions for the acquisition and estimation of the surface appearance of a real object from video sequences, acquired in general and fixed lighting condition. The reconstructed appearance is projected over the high quality 3D geometry of the same object, acquired for example with 3D scanning techniques. The common framework of these new techniques is composed by two stages:

- the registration of the video sequences over the 3D model of the artwork of interest by calibration of the camera; for each video-frame it recovers the camera parameters that permit the projection of a 3D word point into the image plane and then to associate to the 3D point the color of the corresponding pixel in the image;

- the reconstruction of the surface appearance with different degree of quality, from a simple view-independent representation, like the perceived diffuse color, to a more complex one, like the Surface Light Fields or the Spatially Varying Bidirectional Reflectance Distribution Function (SV-BRDF).

The objectives to reach for new proposed techniques are:

- a flexible and general method that can be adapted to more general cases characterized by objects with different shape, size and surface appearance;

- an easy acquisition in the natural environment of the object without using lab lighting conditions;

- an automatic system to reduce the time for the acquisition and the calibration of the data;

- an automatic system to reduce the expertise of the operator, needed to evaluate if the acquired data are sufficient for the reconstruction of the appearance.

The proposed approach takes advantage of the two main differences of a video sequence with respect to a traditional photographic acquisition. The first one is the high temporal coherence and the redundancy of the data, which can be used to improve the visual quality of the appearance reconstruction by using techniques that deal with shadow, shading and highlights. The second one is the easy of the acquisition, which allows the capture of many views of the object in a short time. The positive consequence is an easier capture since the operator is facilitated in acquiring a complete sampling of the surface reflection.

1.1.1 Not only 3D

Reflectance Transformation Imaging (RTI) techniques are an example of consolidate and wide used technology for the acquisition of the surface appearance in the Cultural Heritage Field. RTIs are image-based methods that allow the acquisition of the Surface Reflectance Field of an object. Starting from a set of photos acquired from a fixed viewpoint under varying lighting conditions (a single predominant directional light that is moved in front of the object), RTI encodes the Surface Reflectance Field in a compact way, using a view-dependent per-pixel reflectance function. This function enables the interactive re-lighting of the object from any light direction. RTI techniques are widely used in the virtual examination and study of Cultural Heritage artworks, like documentation tools and to support detailed visual analysis, giving a precious instrument to the specialists in the interpretation process. Their success and wide use in the Cultural Heritage field are due to several reasons: inexpensive and widely available hardware (in many cases just a digital camera and a light); scale well to both large and very small objects; it is easy to achieve such high sampling density and precision that most current 3D scanners are unable to reach, even under optimal acquisition conditions; it is possible to obtain optimal reproduction of the reflectance of materials, like gold, which are challenging to model and acquire with other methods.

The main drawback of this technology is the partial approximation of the surface appearance that does not allow the changing of the viewpoint. This reason reduces the applicability of the RTI technology to the objects with predominant flat geometry, where a single viewpoint is sufficient to acquire the majority of the most interesting features, like paintings, coins, bas-reliefs and inscriptions.

In this domain, there exist two important issues to address: how to provide better and more flexible virtual inspection capabilities with a set of operators that improve the perception of details, features and overall shape of the artwork; how to increase the possibility to disseminate this data and to support remote visual inspection of both scholar and ordinary public.

Therefore a parallel research activity, presented in this thesis, is the proposal

of new solutions to address these open questions in RTI field. The first solution introduces a new shading enhancement operator that creates a locally adaptive multiple-light illumination environment, improving the RTI image sharpness and illumination of the surface features. The basic idea is to combine, in a single view, the contributions of lights coming from different directions, such that different portions of the image are illuminated in a not-physically consistent but appealing way. The second solution presents an interactive visualization system, designed to be deployed either locally by a museum kiosk or remotely by a web site, for the real-time presentation and virtual inspection of RTI images using the HTML5 and the WebGL technologies.

1.2 Contribution

The main contributions, presented in this thesis, concern several important open issues in Computer Graphics and Computer Vision:

- **Video-to-Geometry and Image-to-Geometry Registration**. A fundamental step to transfer the color information from the images to the 3D model is the camera registration procedure, which allows the calculation of the camera parameters that define the projection process of a 3D point into the image plane. An innovative solution for the accurate and efficient alignment of a video sequence of a real object over its 3D model is proposed. The algorithm takes advantage from the temporal coherence and the redundancy of the video data, putting together two different Image-to-Geometry registration approaches: feature-based registration by KLT video tracking; statistic-based registration by maximizing the Mutual Information between the gradient of the frame and the gradient of the rendering of the 3D model with some illumination related properties, such as the surface normal and the ambient occlusion. While KLT tracking speeds up the registration, allowing a fast alignment of short sequences with simple camera movements, and steers the convergence of Mutual Information towards good camera parameters, the Mutual Information corrects the drifting effect that the KLT tracking produces over long sequences, due to the incremental tracking and the camera motion.

- **Diffuse Color Estimation**. The first stage of the proposed methods for the surface appearance estimation is the computation of the diffuse color component. Two similar approaches are proposed. The key idea is to try to reconstruct the acquisition lighting environments in order to make the computation of the diffuse color more robust and accurate by means of a weighted approach. The acquisition environment is approximate with the creation of an environment map by means of specular mirror reflection of the color samples with a higher probability to show a not diffuse reflectance behavior. The first proposed method is based on few fixed but general luminance thresholds that

decrease the statistical influence of the samples with a not diffuse behavior in the computation of the final color. The second approach is based on a statistical framework that makes the method more automatic and robust.

- **Surface Light Field Reconstruction**. Surface Light Field is a partial approximation of the surface appearance that allows changing only the viewpoint taking fixed the lighting condition. To generate artifact free renderings, many state of the art algorithms require a dense and uniform sampling of the view direction, obtained for example with special devices like camera arrays or robotic arms. In order to make this technology usable in the Cultural Heritage domain, a new automatic method is proposed. It is able to estimate the Surface Light Field from partial and irregular video acquisition, made by simple movements around the object, reducing the expertise of the operator needed to evaluate if the acquired data are enough for the reconstruction. The main idea is to separate the estimation of two components: the diffuse color, using statistical operations, and the other residual Surface Light Field effects, using a linear combination of spherical functions. In this way, the proposed solution avoids the occurrence of rendering artifacts, like ringing and banding effects, normally caused by the irregular viewpoint sampling and the fitting of the color in the basis of spherical functions.

- **Spatially Varying BRDF Estimation**. The reconstruction of the Spatially Varying BRDF allows renderings of the object in new virtual environments with a complete interaction with both the light and the view direction. The state of the art solutions are difficult to adapt to the Cultural Heritage field, which needs an easy acquisition system that should reduce the expertise of the operator and should be able to work with different types of objects and materials in few controllable lighting conditions. In order to simplify the acquisition phase and to obtain at the same time accurate and photo-realistic renderings of the object, an innovative statistical approximation method is proposed. The new algorithm estimates the SVBRDF starting from video sequences acquired in general and fixed lighting conditions. It passes through the reconstruction of the acquisition environment maps and the estimation of the main appearance components: the diffuse and the specular reflection. The trade-off between the easy of acquisition and the obtained results makes the algorithm useful for practical applications.

- **RTI Shading Enhancement**. One of the main applications of the RTI images is the study and analysis of Cultural Heritage artworks. In order to make these virtual studies more powerful, some shading enhancement algorithms have been developed to improve the readability and the interpretation of the most interesting details depicted on the surface of the artwork. Since the available methods do not explore all the potentialities of the data store in an RTI image, a new shading enhancement method is presented. The method, called

Multi-Lighting Detail Enhancement, combines, in a single view, the contributions of different lights coming from different directions, such that different portions of the image are illuminated in a not-physically consistent, but appealing way, improving the perception of details, features, and overall shape of the artwork.

- **RTI Web Visualization**. The recent advances of the web visualization instruments are increasing the capability to disseminate and support the remote visual inspection of interactive 3D content, making easier the interaction with the 3D data directly in the web browser without external plugings. In order to take advantage from this innovation, an interactive visualization system is proposed for the real-time presentation and virtual inspection of RTI images using the HTML5 and the WebGL technologies. The viewer was also recently tested in a concrete project: the development of a interactive kiosk to virtually present the coin collection of the National Museum of San Matteo in Pisa, designed to be deployed either locally (on a touch screen in the museum) or remotely (by a web site).

1.3 Thesis Structure

The thesis is organized as follows.

Chapter 2 provides a general overview about the state of the art of the research areas involved in the main topics of the thesis. The first part presents the camera calibration problems and the different methods proposed to reach the optimal image-to-geometry registration. The second part introduces the structure from motion methods and their application for the 3D reconstruction and camera calibration from set of images and videos. The third part is dedicated to the estimation of the light scattering function from calibrated images and its simplified version for opaque objects, which do not exhibit complex reflectance behavior.

Chapter 3 presents a new algorithm for the accurate and efficient alignment of a video sequence of a real object over its dense triangular mesh. The solution uses two different Image-to-Geometry registration approaches: feature-based registration by KLT video tracking and statistic-based registration by maximizing the Mutual Information between the frame and the rendering of the 3D model. The effectiveness and performance of the algorithm are tested on a synthetic sequence with known camera parameters, in order to evaluate the maximum registration error, and on four real sequences of objects with different features. Then the chapter presents a comparative study to evaluate the performance improvements obtained using the gradient maps in the registration by Mutual Information with respect to the original algorithm that was extended.

Chapter 4 presents a new method for the estimation of the Surface Light Field using video sequences with an irregular and not uniform sampling density of the

viewpoint. The method is based on the separation of the diffuse color and the residual reflectance effects. The chapter describes the three main steps of the algorithm: the estimation of the direction of the main light sources, which were in the acquisition environment; the estimation of the diffuse color by reducing the statistical influence of the color samples with a higher probability to have a not diffuse behavior; the fitting of the residual component in a linear combination of spherical functions. The rendering results obtained for three different objects are shown.

Chapter 5 presents an innovative algorithm for the statistical approximation of the Spatially Varying Bidirectional Reflectance Distribution Function acquired in general and fixed lighting conditions. Starting from the alignment of some videos on the 3D model of the object, the chapter introduces the main phases of the algorithm: the approximation of the environment map of the acquisition scene, using the same object as a probe; the estimation of the diffuse color of the object with statistical operation; the estimation of the specular components of the main materials of the object, starting from a partial user-assisted material segmentation. The obtained rendering results on four different test cases are finally discussed.

Chapter 6 deals with RTI technologies: shading enhancement and interactive web visualization. The first part presents a new shading enhancement algorithm that takes advantage from the possibility to use different light direction in different areas of the image. The second part presents a new interactive presentation system for multimedia data and RTI images, developed in HTML5 and WebGL.

Finally, Chapter 7 analyzes the proposed solutions, providing the list of publications produced by the main contributions of the thesis, and proposes possible future extensions.

Chapter 2

Background and Related Work

The acquisition of the surface appearance, together with the reconstruction of the 3D geometry, is a fundamental pre-requisite for producing photo-realistic renderings of real objects in several application fields. The estimation of the appearance is a complex task that can be achieved with two different approaches: to estimate at the same time geometry and reflectance from a set of images; to split the reconstruction of geometry and reflectance, starting from a previously acquired high quality and high resolution 3D model of the object. In the last case, which is the approach adopted by the new techniques proposed in this thesis, two steps are required: the registration of a set of photos over the 3D geometry; the estimation of the chosen appearance model using the color projected from the aligned images.
This chapter presents an extensive overview of the state of the art in three important research fields: the camera calibration and the image-to-geometry registration (Section 2.1); the structure from motion solutions (Section 2.2) and their applications for 3D reconstruction and camera calibration from set of images, with a special attention to the use of their output data to solve the image/video-to-geometry registration problem; the estimation and approximation of the light scattering function of the object materials from calibrated images, limited to opaque objects that do not exhibit complex reflectance effects like sub-surface scattering, phosphorescence and fluorescence (Section 2.3).

2.1 Camera Calibration

Camera calibration is the process necessary to find the camera parameters that define how the 3D world points are projected into the image plane. This process is the basic step for the more complex image-to-geometry registration procedures that require the alignment of one or more images of the same object, taken at different times and from different viewpoint, to its 3D geometry. The fundamental aspects of the camera calibration process are: the camera model that describes the imaging process, typically a perspective camera; the type of data, to give in input

at the optimization procedure for the computation of the camera parameters and to constraint the relation between the image and the 3D model, which can be explicit, using a set of 2D/3D correspondences, or implicit, by statistical correlation between the image and a special rendering of the 3D model.

2.1.1 Perspective Camera Model

The perspective camera is defined by two groups of parameters:

- *intrinsic parameters*, which are related to the internal characteristics of the camera;
- *extrinsic parameters*, which are associated with the position and the orientation of the camera in the space.

These parameters give the possibility to project any point in the 3D space in the corresponding point on the image plane of the camera. In fact, given the camera model (see Figure 2.1), the 3D point M expressed in the Euclidean world coordinate system W_c and the 2D point m in the image coordinate system $(\vec{u}, \vec{v})$, the projection can be described by the following equation:

$$m' \simeq KR[I|t]M' \tag{2.1}$$

where m' and M' are the homogeneous coordinate of m and M and $\simeq$ means equal up to a non-zero scale factor.

The 3×3 matrix K depends on the intrinsic parameters of the camera and represents the transformation from a point in the camera coordinate system to a homogeneous point in the image plane. It can be written as:

$$K = \begin{bmatrix} \frac{f}{k_u} & s & u_0 \\ 0 & \frac{f}{k_v} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.2}$$

where f is the focal length of the camera in millimeters, which is the distance between the optical camera center and the image plane, (k_u, k_v) are the dimensions in millimeters of each pixel of the image plane, s is the skew parameter for the axes $(\vec{u}, \vec{v})$ of the image plane (equal to zero for orthogonal axes), (u_0, v_0) are the coordinates of the principal point, which is the intersection between the optical axis of the camera and the image plane.

The 3×4 matrix $R[I|t]$ depends on the extrinsic parameters and represents the Euclidean transformation of a homogeneous point M' from the world coordinate system W_c to the camera coordinate system C_c. R is a rotation matrix that represents the orientation of the camera coordinate system and the vector t is a translation from the origin of the world to the origin of the camera.

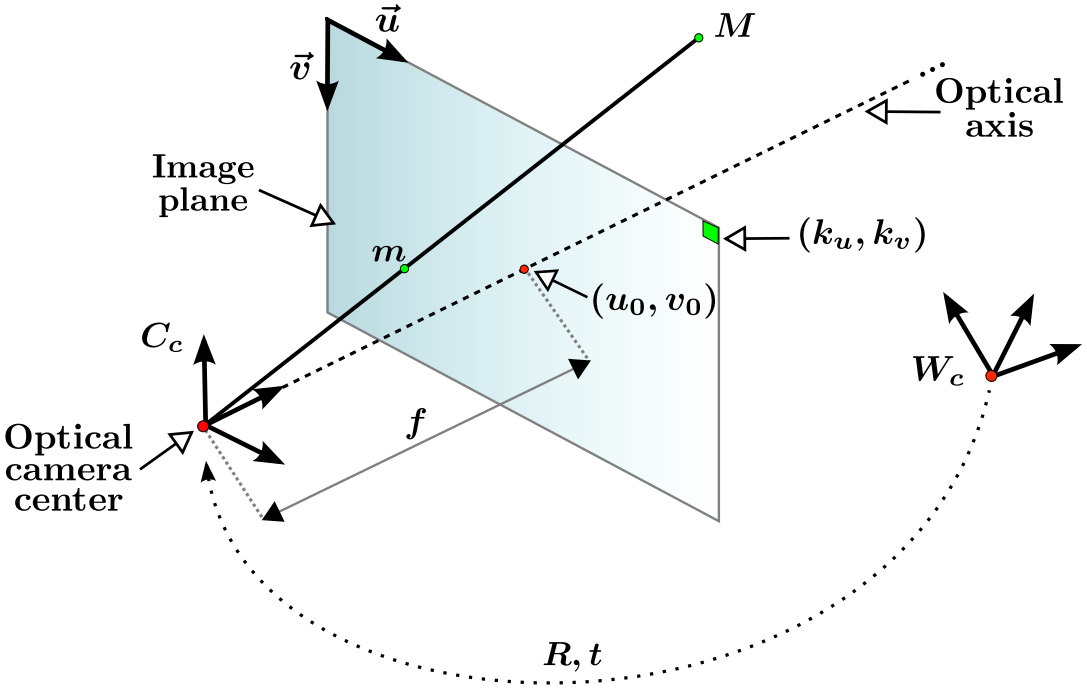


Figure 2.1: *Perspective camera model.*

Another aspect of the camera that must be taken into account is the distortion introduced by the lens. There exist two types of distortions: radial distortion, mainly caused by flawed radial curvature of the lens, and decentering distortion, due to the optical center of the lens that is not correctly aligned with the center of the camera. The total distortion is the sum of these two distortions:

$$\delta_x = \delta_{xr} + \delta_{xd} \qquad \delta_y = \delta_{yr} + \delta_{yd} \tag{2.3}$$

and it can be added to the distorted 2D point $m_d = (x_d, y_d)$ to obtain the corresponding undistorted one $m_u = (x_u, y_u)$:

$$m_u = m_d + \begin{pmatrix} \delta_x \\ \delta_y \end{pmatrix} \tag{2.4}$$

The two distortions are computed as follow:

$$\begin{aligned}
\delta_{xr} &= x_d(k_1 r^2 + k_2 r^4 + \dots) && (2.5)\\
\delta_{yr} &= y_d(k_1 r^2 + k_2 r^4 + \dots) && (2.6)\\
\delta_{xd} &= 2t_1 x_d y_d + t_2(r^2 + 2x_d^2) && (2.7)\\
\delta_{xd} &= t_1(r^2 + 2y_d^2) + 2t_2 x_d y_d && (2.8)\\
r &= m_{d\ 2} && (2.9)
\end{aligned}$$

where $k1, k2, \dots$ are the radial distortion coefficients (two coefficients are enough to model the distortion), and t_1, t_2 are the decentering distortion coefficients. Often the decentering distortion is neglected, because its influence is not very significant.

2.1.2 Camera Calibration Methods

The basic idea of the camera calibration process is to estimate all the parameters of the camera model, 6 extrinsic parameters (3 for the rotation, if R is parameterized by the Euler angles, and 3 for the translation) and 8 intrinsic parameters (focal length, pixel dimensions, principal point, skew, and two radial distortion coefficients). Even if several methods were proposed to estimate all the parameters, usually the problem is simplified by assuming some intrinsics provided by the camera manufactured. Generally, the calibration methods adopt solutions based on the minimization of a non-linear error function by using a non-linear least-square minimization algorithm. The most used is the Levenberg-Marquardt algorithm [110] that puts together the advantages of the Gauss-Newton algorithm and of the Gradient Descent method.

The Direct Linear Transformation (DLT) algorithm [1] is the simplest approach proposed to estimate all elements of the whole projection matrix $P = KR[I|t]$. This method, which does not estimate the lens distortion coefficients, extracts for each correspondence between the 3D point $M = (X, Y, Z)$ and the 2D point $m = (x, y)$ two linearly independent equations:

$$p_{11}X + p_{12}Y + p_{13}Z + p_{14} - p_{31}Xx - p_{32}Yx - p_{33}Zx - p_{34}x = 0 \tag{2.10}$$

$$p_{21}X + p_{22}Y + p_{23}Z + p_{24} - p_{31}Xy - p_{32}Yy - p_{33}Zy - p_{34}y = 0 \tag{2.11}$$

If p is a vector of all coefficients of P, the equations can be rewritten in the form $Ap = 0$, where A is a $2n \times 12$ matrix and n the number of correspondences between 3D coordinates and 2D image points. Now the elements of the projection matrix P can be computed by using the Singular Value Decomposition of A. At least $n = 6$ correspondences are necessary to solve the system of equations. The intrinsic and extrinsic parameters can be extracted from P by using a QR decomposition. The results of DLT method are often used as an initial guess for other calibration methods.

An approach similar to DLT is proposed by Faugeras and Toscani [56] to improve the extraction of intrinsic and extrinsic parameters from the matrix P. This technique is further improved [52] using a non-linear method to minimize an error function defined as the distance between the image point and the projection obtained by the estimated camera. It requires at least 6 correspondences and performs optimization on extrinsic parameters and the focal length value.

Tsai's calibration model [192] is a two-step method that models the radial distortion assuming that some camera parameters are provided by the manufacturer. This assumption reduces the number of parameters in the first step where an initial guest is estimated. It requires at least 8 2D-3D correspondences. In the first step, the method computes an initial guess of the orientation and position of the camera. In the second step, it computes the focal length and the distortion coefficients. Finally, it executes a non-linear optimization step.

Another two step method is proposed by Heikkila and Silven [88]. In the first step, it extracts an initial estimation of the camera parameters using a closed-form

solution (for example DLT). Then a nonlinear least-squares estimation, employing the Levenberg-Marquardt algorithm, is applied to refine the output and to compute the distortion parameters. The model uses two coefficients for both radial and decentering distortion, and the method works with single or multiple images and with set of coplanar or non-coplanar 3D points.

Zhang's calibration method [214] requires a planar checkerboard grid to be placed at different orientations (more than 2) in front of the camera. The developed algorithm uses the extracted corner points of the checkerboard pattern to compute a projective transformation between the image points of the n different images, up to a scale factor. Afterward, the intrinsic and extrinsic parameters are recovered using a closed-form solution, while the radial distortion terms are recovered within a linear least-squares solution. A final nonlinear minimization of the reprojection error refines all the recovered parameters.

An interesting overview and comparison of the camera calibration methods can be found in [172] and [163].

2.1.3 Camera Pose Estimation

If the intrinsic parameters are known, the calibration process is reduced to estimate only the extrinsic camera parameters. With a given set of n correspondences between 3D world coordinates and 2D image points, the six degrees of freedom of the camera pose can be estimated. This problem is often referred to as the Perspective-n-Point (PnP) Problem. It is also possible to use the DLT algorithm [1] for estimating only the extrinsic parameters by simply multiplying the estimated matrix P with K^{-1}, but the results are not very stable. The problem of estimating the camera pose has been extensively studied in the literature. The methods can be classified into two categories: iterative and non-iterative approaches. The non-iterative methods are often used to estimate the pose without any prior knowledge, while purely iterative methods need a first guess of the extrinsic camera parameters.

All the iterative methods usually define an error function depending on a given camera pose and minimize these error functions iteratively. The error function can be defined either in image space or in object space. A very popular solution was presented by Davis et al. [42]. The method, called POSIT, computes an approximate solution, by solving a linear system using the scaled orthographic projection model, and then the camera pose is iteratively refined. A problem of this approach is that it cannot be applied when the points are coplanar. In [148] a similar approach is described, which handles the coplanar case. SoftPOSIT [40] is another interesting method that handles the extrinsic camera parameters estimation and the determination of the correspondences. This can be useful for problems where the connection between 3D points and 2D points is ambiguous.

The non-iterative approaches rely on a first estimate of the depth and 3D positions of a feature point in the camera coordinate system. Then the rotation R and translation t can be easily obtained by aligning the points with a closed-form

solution [93]. Non-iterative methods usually have a high complexity. To overcome this problem a very efficient and accurate non-iterative algorithm was developed by Moreno et al.[147]. The central idea is to express the n 3D points as a weighted sum of four virtual control points and to solve the system in terms of their coordinates. Thereby the complexity is reduced to $O(n)$.

Usually non-iterative methods are employed to compute an initial estimate of the camera pose, while iterative methods are more accurate and are used to refine the estimation result.

2.1.4 Robust Estimation

In some cases, the manual or automatic process that selects the 2D/3D correspondences to use in the camera calibration procedure can return ambiguous or inaccurate correspondences. These spurious measurements will have a great influence on the estimated camera pose. Therefore, a detection of incorrect measurements is indispensable for robust pose estimation.

A method for robust estimation is RANSAC (RANdom SAmple Consensus), presented by Fischler and Bolles [59]. From an observed set of data, a smallest possible subset of samples is randomly selected and used to estimate the model parameters. Then the other samples are tested to verify the number of them that fit to the model. For robust pose estimation, this means that the minimum required number of correspondences is selected to estimate a camera pose. All other 3D points are then projected with that camera pose into the image plane, and it is tested how many points exist which have a smaller re-projection error than a certain threshold. Such correspondences are called inliers. The other correspondences with a big error are called outliers. If the amount of inliers is not big enough, a camera pose is estimated with another random subset of correspondences. This process is iterated until the amount of inliers exceeds a threshold or if a maximum number of iterations is reached. If the RANSAC method has been applied successfully, the pose can be refined by applying a non-linear method on all inliers.

RANSAC has received many improvements: MSAC and MLESAC [189], which choose the solution to maximize the likelihood rather than just the number of inliers; Locally Optimized RANSAC (LO-RANSAC) [29], which introduces some local optimization methods to apply when a new maximum in the number of inliers has occurred; PROgressive ranSAC (PROSAC) [28], based on an ordering of the set of initial samples in order to do a semi-random selection of the first input data; RANSAC for quasi degenerate data [62].

2.1.5 Semi-Automatic Image-to-Geometry Registration

The image-to-geometry registration allows the alignment of one or more images of the same object take at different times and from different viewpoint to a previous acquired geometry, using for example 3D scanning techniques.

A robust semi-automatic approach was proposed by Franken et al. [63] for general cases. A tool allows the user to set correspondences between the 3D model and an image or between images. The main contribution is a technique to minimize the user intervention. The main idea is to setup a graph of correspondences, where the 3D model and all the images are represented as nodes and two nodes are connected if a correspondence between them exists. The graph of correspondences is then used to automatically infer new 2D/3D correspondences and to find the shortest path, in term of the number of correspondences that must be provided by the user, to complete the registration of all images (Figure 2.2).

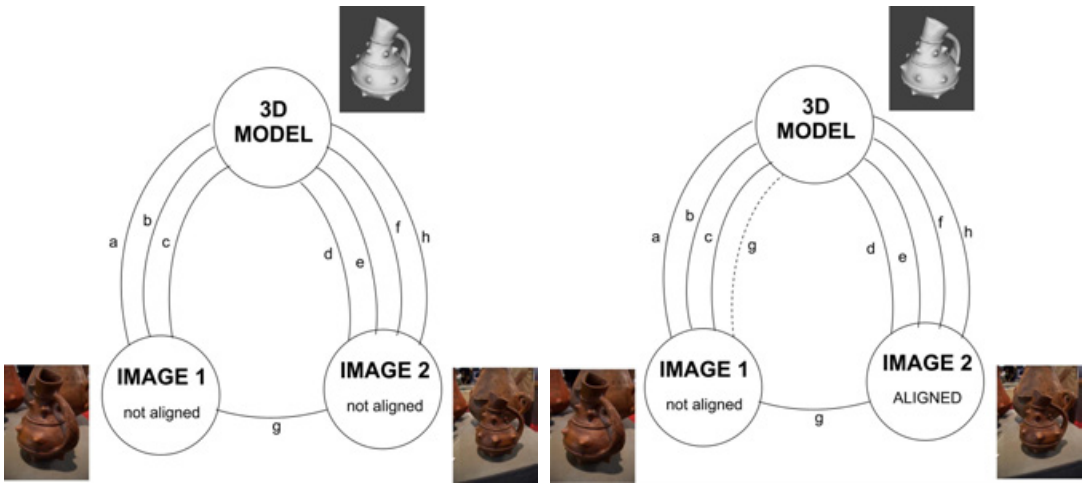


Figure 2.2: *Example of correspondences graph from [63]. A new 2D/3D correspondence (g) for IMAGE1 is inferred automatically given an image-to-image correspondence which links IMAGE1 with IMAGE2.*

An automatic planning to select the minimal set of camera position to cover the entire surface could lead to good results [129] and reduce the importance of registration, but in most cases no information about camera positions is known in advance.

2.1.6 Automatic Image-to-Geometry Registration

The problem of automatically aligning a set of uncalibrated images to a 3D model is important both in Computer Graphics and in Computer Vision.

Automatic registration can be achieved by analyzing the image features [141] or using the reflectance value acquired during scanning [94]. Neugebauer et al. [141] present a hybrid approach where the estimation based on correspondences is combined with a registration based on the analysis of the image features, like the edge intensity. This semi-automatic approach needs a preliminary calibration of the intrinsics of the camera. Moreover, one of the hypotheses is that the illumination must be the same for all the images. Liu et al. [115] propose a novel algorithm

for the 3D range scan to 2D image registration problem in urban scene settings. Assuming that the 3D scene contains clusters of vertical and horizontal lines, they used orthogonality constrains for the registration. In the specific parallelepipeds are extracted from the range maps, and subsequently matched to the rectangle extracted from the input images. Larue et al. [104] propose a hardware setup based on structured light 3D scanner and camera for the automatic registration and calibration of images and 3D data needed to estimate a Surface Light Field. The main idea is to project a parameterization over the surface in order to make automatic the finding of the point-to-point and the point-to-pixel correspondences.

Several papers rely on the analysis of the silhouette of the object [21][107]. These methods try to find the camera transformation by minimizing the error between the contour found in the image and the contour of the projected 3D model. The error is typically computed as sum of distances between sample points on one contour to the nearest points on the other [119]. Another approach computes the sum of minimal distances of rays from the eye point through the image contour to the model surface, which are computed using 3D distance maps [21]. The work by Lensch et al. [107] proposes a robust implementation of previous silhouette based techniques, introducing a similarity measure to compare them (Figure 2.3). Unfortunately, the use of silhouette matching has two important limitations: it must be easy to distinguish the object with respect to the background; the object must be entirely inside each image. This can be a very important drawback when a big object must be acquired preserving fine color details.

Another class of methods is based on image similarity measures, like the Mutual Information used in the multi-modal image registration. The Mutual Information (MI) is a measure of statistical dependency between two datasets and it is particular suitable for registration of images acquired with different modalities. From an information theory viewpoint, given two random variable A and B, the Mutual Information is the amount of information about B that A contains.The first methods proposing this technique were developed by Viola and Wells [199] and by Maes et al. [125]. The Viola's alignment approach uses the mutual information between the surface normal and the image brightness to correlate shading variations on the image with the surface of the model (Figure 2.4). Leventon et al. [111] extended this alignment framework to use multiple views of the object when a single image does not provide enough information.

Since then, several registration methods based on MI have been proposed, especially for medical images [156]. Most of these studies regard simple geometric transformations such as 2D roto-translations or affine transformations. This means that some issues related to the camera model registration are not addressed. Moreover, the resolution of medical data is often quite poor, so using MI in a general case is difficult if no specific adjustments are made. Another key issue in the use of MI is the choice of the optimization strategy to achieve the maximization; the pros and cons of several methods are presented in [126]. An interesting method for 3D object tracking using MI, which allows almost real-time tracking of simple template-based

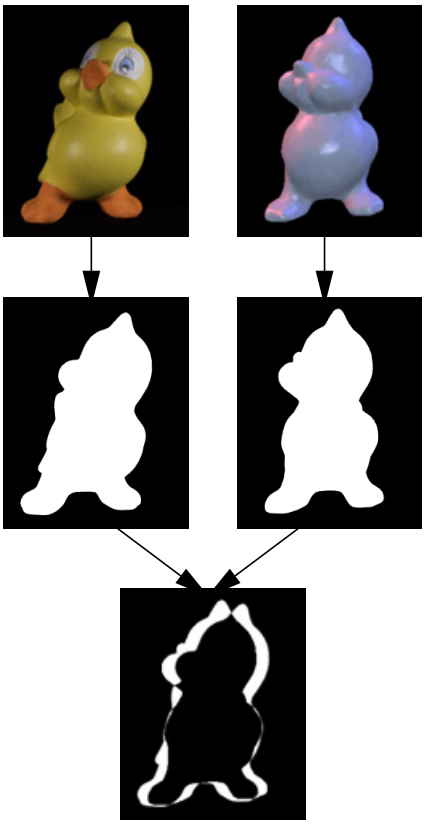


Figure 2.3: *Silhouette comparison from [107]. The silhouettes of image and model are compared to calculate a similarity measure.*

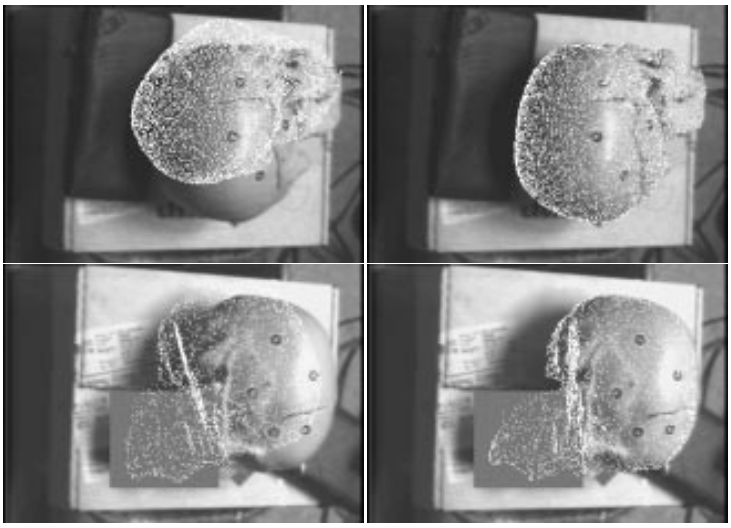


Figure 2.4: *Registration by MI proposed in [199].*

objects, was recently proposed [151].

Regarding more complex texture registration tasks, a system has been developed to improve texture registration by exploiting 2D-2D and 2D-3D MI maximization [30]. However, the optimization is only introduced in 2D-2D registration, while for 2D-3D alignment, Viola and Wells's approach is used.

A more recent approach was proposed in [34], where Viola's approach is extended using several types of rendering with some illumination related geometric properties, such as surface normals, ambient occlusion, specular reflection directions and combined versions of them (Figure 2.5). A new optimization strategy based on the algorithm NEWUOA [160] is used.

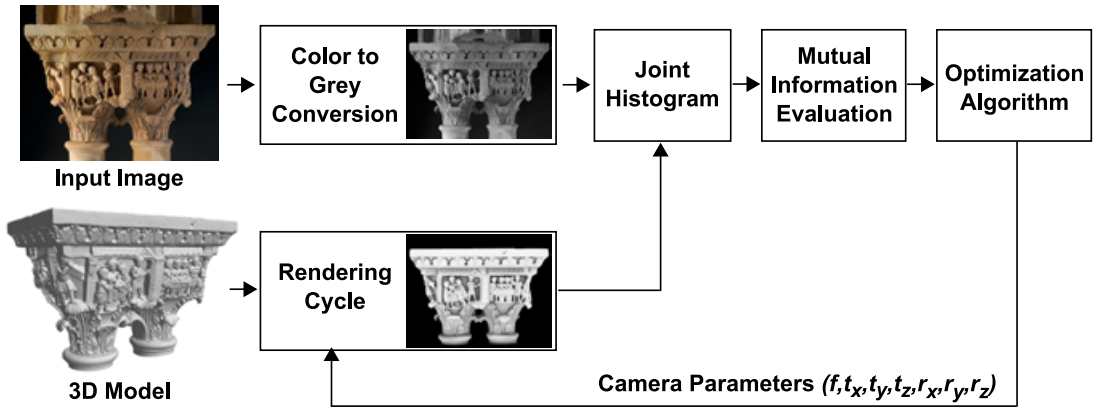


Figure 2.5: *Registration by MI exploiting illumination-related geometric properties proposed in [34].*

2.2 Structure from Motion

The Structure from Motion (SfM) is the process that, given a set of images, recovers simultaneously the 3D structure of the scene and the camera projection matrix using only corresponding 2D features in each image. More formally, given n projected points x_{ij} in m images, for $i \in 1...m$ and $j \in 1...n$, the goal is to estimate a consistent structure of 3D points $X_1, ..., X_n$ and the projection matrix $P_1, ..., P_m$ that allow the projection of the 3D point X_j in the corresponding 2D features x_{ij}. The three main step of the algorithm are:

- Feature detection and matching, to detect and match the most salient 2D features in the images;
- Structure and motion recovery, to compute an initial estimation of the 3D points position and the camera projection matrices;
- Bundle Adjustment, to refine further the estimation of the previous step reducing the reconstruction error.

Structure from Motion is used as an important step in several application fields: 3D Model Reconstruction, for the production of sparse or dense point clouds; 3D Motion Matching, for the automatic reconstruction of virtual reality models from video sequences and for the determination of camera motion so that computer-generated objects can be inserted into video; Camera Calibration and Image-to-Geometry Registration for the automatic or semi-automatic alignment of a set of images over an existing 3D model.

The Structure from Motion solutions are intensively analyzed in this section because their data are partial used in the development of the new method proposed in the Chapter 3 to solve the the video-to-geometry registration problem.

2.2.1 Feature Detection and Matching

The goal is to find and match the same 2D features in the different images of the input set. Two categories of solutions have been proposed. Some solutions adopt marker-based tracking, where artificially designed markers, easy to detect with image processing algorithms, are used to simplify the detection and the creation of 2D correspondences [100][165][140]. Even if the detection and tracking of markers are very reliable, in some cases the preparation of the scene with them is not possible. In such cases, a marker-less solution, based on the natural features of the environment, must be used.

For a marker-less solution, two important elements are needed: a feature detector, which extracts the most salient 2D features of the image, usually points or lines; a feature descriptor, which associates to each extracted feature a descriptive information, usually in form of vector, to use in a matching process. Both of them must have some peculiar characteristics. A good detector should be repeatable and reliable. Repeatability means that the same feature must be detected in different images. Reliability means that the detected point should be distinctive enough so that the number of its matching candidates is small. A descriptor should be invariant to rotation, scaling, and affine transformation, so that the same feature on different images will be characterized by almost the same value, and distinctive to reduce the number of possible matches.

Feature Detector

The Harris corner detector [83] is a well-known point detector that is invariant to rotation and partially to intensity change. However, it is not scale invariant. The detector is based on a local auto-correlation function that measures the local changes of the image. For each point (x, y) it computes the Harris matrix:

$$A = \sum_{(u,v)\in W} w(u,v) \begin{bmatrix} I_x^2(u,v) & I_x(u,v)I_y(u,v) \\ I_x(u,v)I_y(u,v) & I_y^2(u,v) \end{bmatrix} \tag{2.12}$$

where (u, v) is a image point belong to the window W around (x, y) (circular weighted window if $w(u, v)$ is a Gaussian), and I_x and I_y are the partial derivatives of the image I. Then the value of the function $R = det(A) + k(tr(A)^2)$ enables the classification of the point as a corner ($R > 0$), a flat region ($R \simeq 0$), or an edge ($R < 0$).

A simple and efficient detector, named SUSAN (Smallest Univalue Segment Assimilating Nucleus), was introduced in [179]. It computes the fraction of pixels within a neighborhood that have similar intensity to the center pixel. Corners can then be localized by thresholding this measure and selecting local minimum. The FAST (Features from Accelerated Segment Test) detector was proposed in [168]. A point is classied as a corner if one can find a sufficiently large set of pixels on a circle of fixed radius around the point such that these pixels are all significantly brighter than the central point.

Scale invariant detectors [120][133] search for features over scale space. Lowe et al. [120] searches for local maxima of difference of Gaussian (DOG) in space and scale. Mikolajczyk et al. [133] use Harris corners to search for features in the spatial domain and then use a Laplacian in scale to select features that are invariant to scale.

An affine invariant detector is defined by Tuytelaars et al. [193]. Starting from a local intensity maximum, it searches along rays through that point to find local intensity extrema. The link formed by those extrema defines an interest region, which is later approximated by an ellipse. By searching along many rays and using ellipses to represent regions, the detected regions are invariant to affine transformation.

Bay et al. [12] proposed a scale-invariant feature detector based on the Hessian-matrix, but rather than using a different measure for selecting the location and the scale, the determinant of the Hessian is used for both. More precisely, they detect blob-like structures at locations where the determinant of the Hessian is maximum. The Hessian matrix is roughly approximated, using a set of box-type filters.

An extensive survey about local feature detection can be found in [194].

Feature Descriptor

One of the most robust and used feature descriptor is SIFT (Scale-invariant feature transform) and its following derivations. The SIFT descriptor [120] is a vector with 128 elements that is computed on the local image gradient. It uses a regular grid 4×4 around the feature and computes for each grid the histogram of the image gradient. The eight bins values of each histogram become the values of the feature descriptor. SIFT is invariant to scale, rotation, changes in illumination, noise and partially to view change.

Several improvements of SIFT have been proposed. In PCA-SIFT [101], Principal Component Analysis techniques are applied on the local patches of the image gradient to reduce the dimension of the descriptor (typically 36 elements). The result is a descriptor more robust to image deformation and more compact that

reduces the time for feature matching. In GLOH (Gradient Location-Orientation Histogram) [132], the descriptor is computed in a log-polar location grid around the feature and its size is reduced by PCA. In [35], a new modification for SIFT is proposed. The orientation histogram is computed on an irregular grid where the patches are partially overlapped. This modification increases the robustness against the scale variation.

The SURF (Speed Up Robust Features) descriptor [12] is partly inspired by SIFT. It relies on the Haar wavelet responses computed for 16 sub-regions centered on the feature. The result is a descriptor of 64 elements as robust as the SIFT but that reduces the time for features computation and matching.

In the last years, several descriptors have been proposed to allow as fast as possible comparison and matching using binary vector. Calonder et at. [25] propose BRIEF (Binary Robust Independent Elementary Features). The descriptor vector is composed by binary comparison of the intensity of 512 pairs of pixels after applying a Gaussian smoothing to reduce the noise sensitivity. The positions of the pixels are pre-selected randomly according to a Gaussian distribution around the patch center. Rublee et al. [169] propose the Oriented Fast and Rotated BRIEF (ORB) descriptor. Their binary descriptor is invariant to rotation and robust to noise. Similarly, Leutenegger et al. [109] propose a binary descriptor invariant to scale and rotation called BRISK (Binary Robust Invariant Scalable Keypoints). It is based on a sampling pattern consisting of points lying on appropriately scaled concentric circles. Each point contributes to many pairs in the descriptor and the pairs are divided in two subsets: short-distance and long-distance. The long-distance subset is used to estimate the direction of the keypoint, while the short-distance subset is used to build the binary descriptor after rotating the sampling pattern. Alahi et al. [6] propose a keypoint binary descriptor inspired by the human visual system, more precisely the retina, named Fast Retina Keypoint (FREAK).

KLT Tracking

One of the most used solution for feature detection and matching in a video sequence is the KLT tracking algorithm [188] [176]. The main idea is to use the typical coherence between consecutive frames of a video sequence to find the displacement between each pair of corresponding points. It extends the local estimation of the optical flow proposed in [122], to track a template patch under an affine transformation model, with the assumption of small brightness changes between consecutive frames.

Starting from the point features extracted with the Harris' algorithm [83], the tracker computes the displacement of each features in the next frame, by shifting a local window around the feature, until the similarity measure between the local windows in the two frames becomes maximum. Because the movement between two consecutive frames is typically very small, the searching in the next frame starts from the same position of the feature in the previous frame. Typical similarity

measures for these patches are Sum of Square Differences (SSD) or Normalized Cross Correlation (NCC).

This type of tracking presents a drift problem due to several causes: image noise; geometric distortion; illumination changes; occlusions; fast camera movements; 3D features that leave the cameras field of view and reappear after in the sequence. Different solutions were proposed to compensate illumination changes [98] and to merge unconnected features track [32][186]. A further extension of KLT tracker was proposed by Dame [37], where the SSD is substituted by MI for the feature matching between consecutive frames.

2.2.2 Structure and Motion Recovery

Given the 2D feature correspondences, this step recovers the structure of the scene and the motion information of the camera in three different phases: the extraction of the geometric constraints among views; the estimation of the projection matrices; the computation of the 3D coordinates of the features via triangulation.

The research in 3D reconstruction from multiple views started with two views. If the intrinsics matrix K are known, the essential matrix E [117] is used to represent the constraints between two normalized views. The matrix E is defined by the simple equation $\hat{x}'^T E \hat{x}_1 = 0$, where $\hat{x}'$ and $\hat{x}_2$ are corresponding points in the two views expressed in homogeneous normalized coordinates (obtained by multiplying the image coordinate with the inverse of matrix K).

The research was later extended to the uncalibrated case. The concept of fundamental matrix F was introduced and well studied by Faugeras [57] and Hartley [85]. The matrix F is the generalization of E and the defining equation is very similar: $x'^T F x = 0$. The difference is that the matrix K is unknown and thus the view coordinates cannot be normalized.

The main concepts of two-view geometry are explained in Figure 2.6. X is a 3D point and x and x' its projections in the two views respectively. C and C' are the two camera centers. The line segment that connects them is called the baseline. The line among X, C, and C' defines a plane called the epipolar plane. l and l' are the epipolar lines of the two projections of X. The projection of the camera centers on the other images, e and e', are named epipoles. The relations among all these elements form the epipolar constraints.

Three-view geometry has also been developed. The geometry constraints are presented by trifocal tensors [84] that capture relation among projections of a line on three views. The trifocal tensor defines a richer set of constraints over images. Unlike the fundamental matrix, which defines a point to line relation with a one-to-many relation, line correspondences defined by trifocal tensors are one-to-one.

In a video sequence, computing the epipolar geometry between two consecutive frame is not well determined, because the camera could not been moved sufficiently. Therefore, it is necessary a process to select the most robust frames in the sequence, called key-frames, for the computation of the epipolar constrains. Several solutions

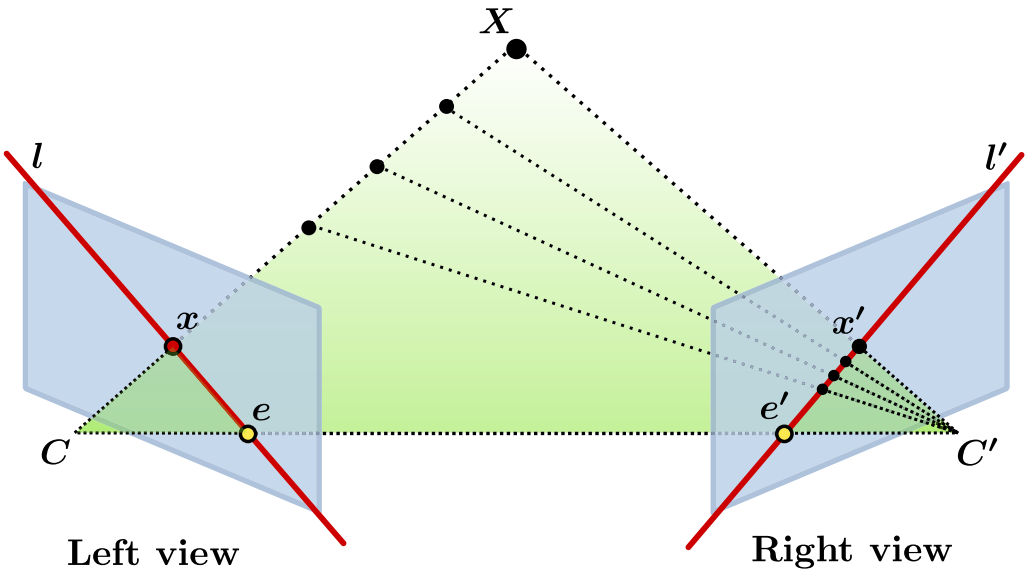


Figure 2.6: *Two-view geometry and epipolar constrains.*

to solve the problem of the key-frames selection based on geometric assumptions are described in [157][76][185].

The following step is to recover the projection matrix from the epipolar constrains. In the calibrated case, the essential matrix can be decompose into the product of a skew-symmetric matrix T, corresponding to the translation, and an orthonormal matrix R, corresponding to the rotation between the views, as proposed in [85]. This decomposition is possible only if the essential matrix has rank 2 and two equal singular values. For this purpose the decomposition can be achieved by computing the singular value decomposition of the essential matrix $E = U\Lambda V$ and by setting the two largest singular values to be equal to their average and the smallest one to zero. The projection matrices follow directly from the recovered translation and rotation by aligning the reference coordinate system with the first camera to obtain:

$$P = K[I \mid 0\,] \qquad P' = K\,[R' \mid T'] \tag{2.13}$$

where T' and R' is defined as follow:

$$T' = U \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} U^T \qquad R' = U \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} V^T \tag{2.14}$$

Four solutions are still possible due to the arbitrary choice of signs for the translation T' and rotation R'. However, the correct one is easily disambiguated by ensuring that the reconstructed points lie in front of the cameras.

In the case of uncalibrated cameras, the camera projection matrices of the two views can be computed from the fundamental matrix using the result showed in

[123]:

$$P = [I \mid 0] \qquad P' = [[e']_x \; F \mid e'] \tag{2.15}$$

where $e' = [e'_1, e'_2, e'_3]$ is the epipole in the right view and $[e']_x$ is a skew-symmetric matrix defined as follow:

$$[e']_x = \begin{bmatrix} 0 & -e'_3 & e'_2 \\ e'_3 & 0 & -e'_1 \\ -e'_2 & e'_1 & 0 \end{bmatrix} \tag{2.16}$$

The projective reconstruction obtained from the fundamental matrix is determined up to an arbitrary projective transformation. Very often, it is necessary to upgrade the reconstruction to a metric one. The solution is the auto-calibration, the process of determining intrinsic camera parameters (matrix K) directly from multiple uncalibrated images. There exist several auto-calibration algorithms based on geometric entities (like absolute conic, absolute dual quadric and absolute dual complex) that are invariant under similarity transformation and with the important feature to be dependent only from the intrinsic matrix K. Solutions have been proposed for the estimation of the intrinsic camera parameters in different conditions: constant parameters for all the views [86][90] and variable parameters[91][158].

The final step is the computation of the 3D point position by triangulation [183].

2.2.3 Bundle Adjustment

From image features x_{ij}, structure from motion gives an initial estimation of the projection matrices P_i and 3D points X_j. But the image measurements are noisy and the equations $x_{ij} = P_i X_j$ will not be satisfied exactly. In this case, the maximum likelihood solution is reached assuming that the measurement noise is Gaussian. The goal is to estimate the projection matrices $\hat{P}_i$ and the 3D point $\hat{X}_j$ that project exactly to the image point $\hat{x}_{ij}$ and also minimize the image distance between the reprojected point $\hat{x}_{ij}$ and the measured image points x_{ij} for every view in which the 3D point appears:

$$\min_{\hat{P}_i, \hat{X}_j} \sum_{i,j} \mathrm{d}(\hat{P}_i \hat{X}_j, x_{ij})^2 \tag{2.17}$$

where the function d is the geometric image distance between homogeneous image points. A more comprehensive treatment of this subject is presented by Triggs et al. [191].

2.2.4 Applications

One of the main applications of the SfM is the reconstruction of sparse 3D point clouds from images. Several pipelines [20][99][197][95] have been proposed to process the images in batch and handle the reconstruction process without making assumptions about the scene.

The key issue is the scalability of the pipeline. One strategy is the partitioning methods that reduce the problem to smaller and better conditioned reconstruction sub-problems, which can be optimized and merged together following [182][144]. Another strategy is to select a subset of input images and feature points that represent the entire solution. Fitzgibbon et al. [60] proposed a hierarchical sub-sampling using a balanced tree of trifocal tensors over a video sequence. In Shum et al. [177] the sequence is divided into segments, which are resolved locally. They are then merged hierarchically using a representative subset of the segment frames. A recent solution [180] that works with sparse datasets describes a method of selecting a subset of images whose reconstruction approximates the result obtained using the entire set. The obtained sparse point cloud is used to propose a novel interface system, called Photo Tourism, for interactively browsing and exploring large unstructured collection of photographs (Figure 2.7). Gherardi et al. [71] proposed a hierarchical and parallelizable scheme for SfM. The images are organized into a hierarchical cluster tree, and the reconstruction proceeds from the leaves to the root. Partial reconstructions correspond to internal nodes, whereas images are stored in the leaves.

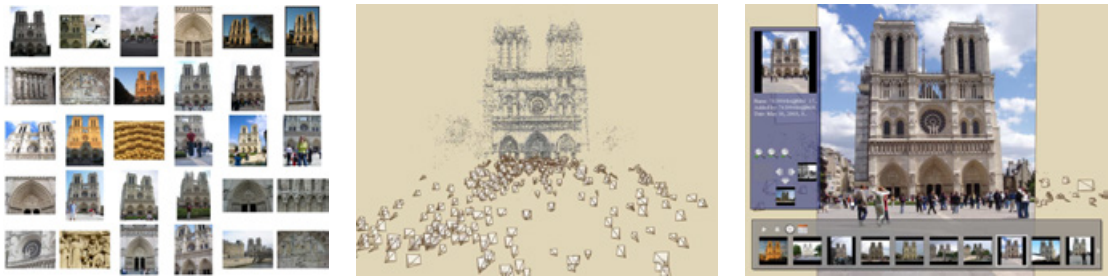


Figure 2.7: *Photo Tourism.*

Another application field of the SfM is the simultaneous alignment of a set of images on a 3D model. Zhao et al. [215] study the registration of a video onto a point cloud. To achieve this goal, a point cloud is computed from the video sequence using motion stereo and camera pose estimation techniques. The obtained point cloud is then registered with the target 3D model using the ICP algorithm. Intrinsic camera parameters must be known beforehand.

Liu et al. [116] present a feature-based method that can be applied under the assumption that the 3D scene contains clusters of vertical and horizontal lines, like urban scenes (Figure 2.8). Given a set of 3D range scans and an independent set of 2D photos, the method generates a pair of 3D models of the scene. The first model consists of a dense 3D point cloud, produced by using a 3D-to-3D registration method. The second model consists of a sparse 3D point cloud, produced by applying a SfM algorithm directly on the image set. The point clouds are automatically aligned with a novel method and integrated with the 2D data in the same reference frame. Stamos et al. [181] extend this system in order to relax the orthogonality

constraint so that the algorithm can be used not only in strictly urban scenes, but also in indoor architectures. A further extension is proposed by Li et al. [114] for indoor environment. The main problem in the indoor environment is the lack of features on large uniform surfaces. The proposed solution uses light projectors to project special light patterns onto the scene surfaces in order to introduce artificial image features.

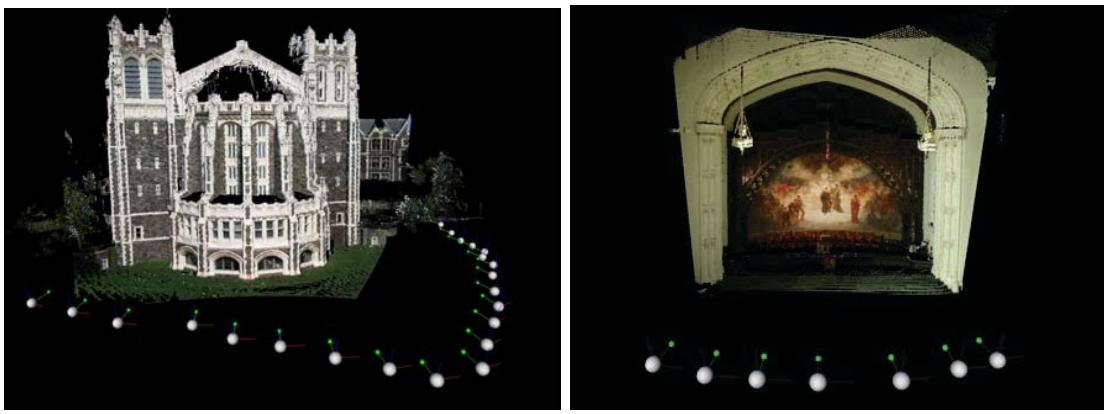


Figure 2.8: *Results of the method proposed in [116] for architectural scene.*

Zheng et al [216] propose a features-based method, which requires the parameterization of the input model in order to extract features based on the surface normal. Corresponding features are extracted in the images that are also calibrated using an SfM algorithm. The matching of 2D and 3D features is done by maximizing a Mutual Information measure.

Pintus et al. [155] propose a method for registering images on point clouds based on three steps: a SfM framework computes the camera parameters and a sparse point cloud; the registration of the sparse point cloud to the given 3D geometry using manual intervention; a specialized sparse bundle adjustment step used to refine intrinsic and extrinsic parameters of each camera.

Corsini et al. [33] present a fully automatic 2D/3D global registration pipeline. The first stage exploits SfM to generate a sparse point cloud from the set of images. During the second stage, this point cloud is aligned to the 3D model using an extension of the 4 Point Congruent Set algorithm for range scan, which takes into account models with different scales and unknown regions of overlap. In the last processing stage a global refinement algorithm, based on Mutual Information, optimizes the color projection of the aligned photos on the 3D object, in order to obtain high quality textures.

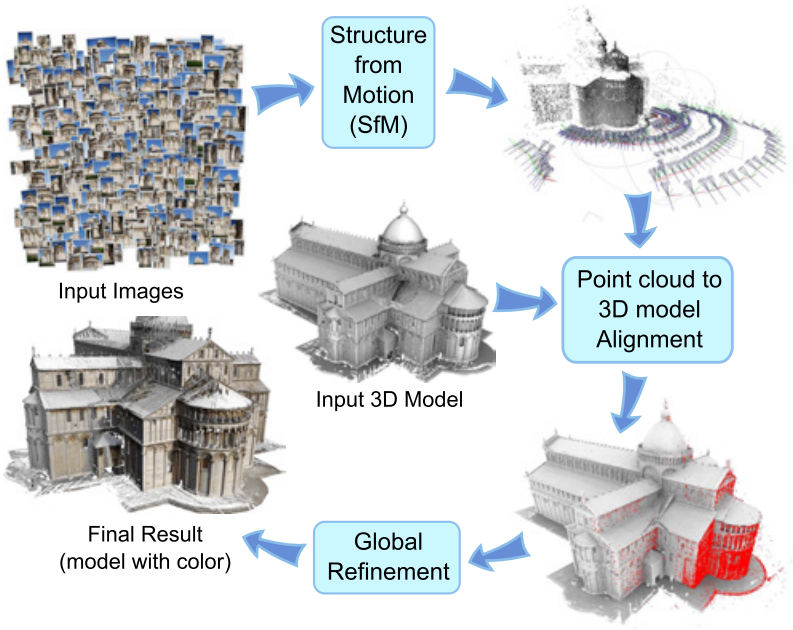


Figure 2.9: *Overview of the global 2D/3D registration pipeline proposed in [33].*

2.3 Surface Appearance

The visual appearance of an object is defined by the way in which it reflects and transmits light [53]. This process is completely described by the light scattering function, a function that involves 12 parameters. The general light scattering process is shown in Figure 2.10. The incident light beam, with direction (θ_i, ϕ_i), hits the surface in the point $\mathbf{X}_i$ (parameterized in 2D space with (u_i, v_i)) at the time t_i with the wavelength λ_i. Now the light can travel through the material and leaves the surface at position $\mathbf{X}_r$ (parameterized in 2D space with (u_r, v_r)), at time t_r, with a changed wavelength λ_r in the direction (θ_r, ϕ_r)

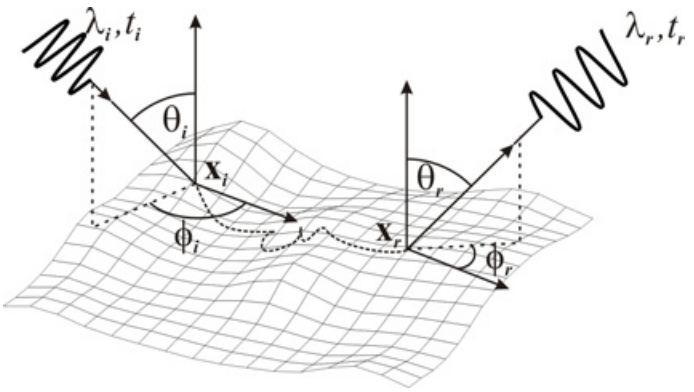


Figure 2.10: *General model of light-material interaction (from [138]).*

Since a mathematical formulation of this function does not exist and the measurement of a 12-dimensional function is currently not practical, several attempts to estimate a simplified version of this function have been proposed. These simplified functions are obtained by introducing constrains on the parameters that reduce the dimensionality of the function and the type of reflectance effects that it is possible to reproduce. A hierarchy of these functions is shown in Figure 2.11. A unified approach for the specification of reflectance in relation to the geometry of both the incident and the reflected light beam and the relative nomenclature were introduce by Nicodemus et al. [145].

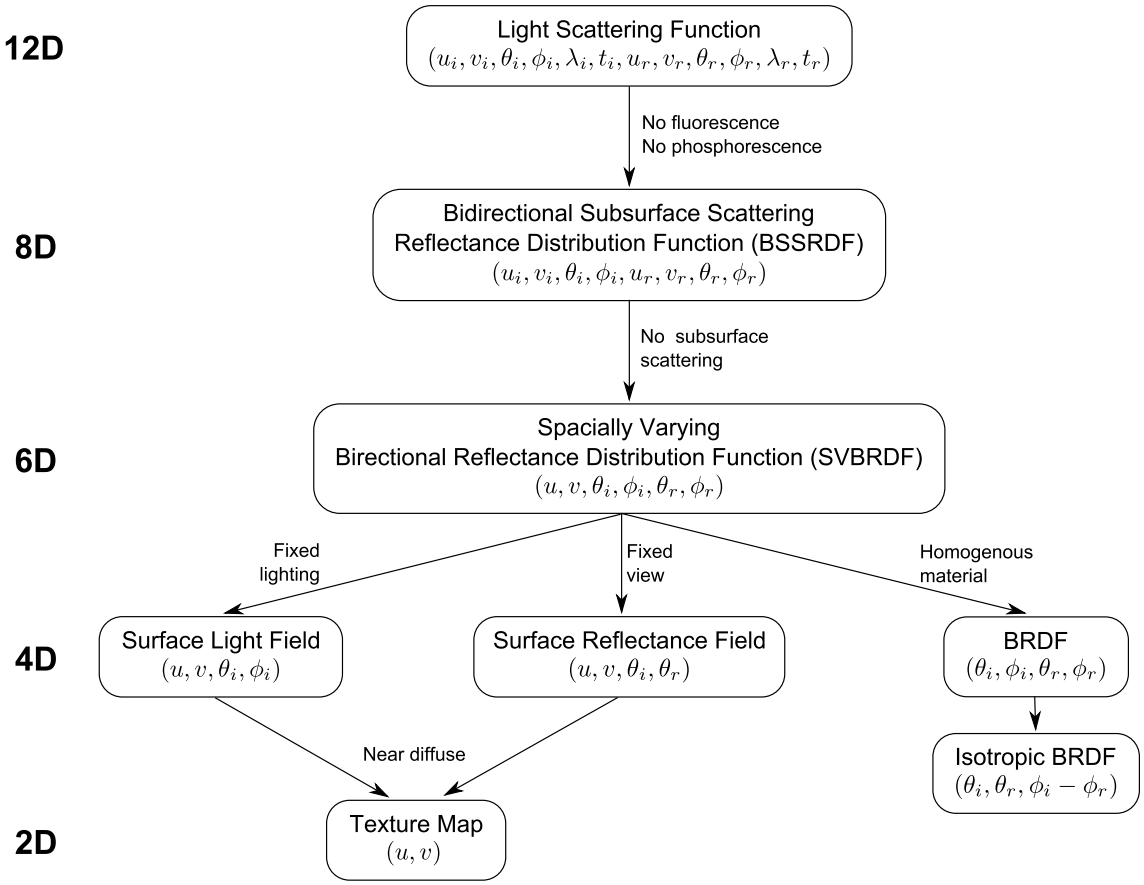


Figure 2.11: *Hierarchy of light scattering functions.*

In the Computer Graphics field, the typical simplification assumptions for the acquisition and rendering of reflectance are: light transport takes zero time ($t_i = t_r$); interaction does not change wavelength ($\lambda_i = \lambda_r$); wavelength is discretized into the three color bands red, green and blue. These assumptions exclude effect like fluorescence and phosphorescence. The final obtained function is called Bidirectional Subsurface Scattering Reflectance Distribution Function (BSSRDF) and it models the process in which light penetrates the surface of a translucent object, it is reflected a number of times inside the material, and it leaves the surface at a different point.

A further simplification can be obtained for opaque objects, where the reflection taking place on an infinitesimal surface element ($\mathbf{X}_i = \mathbf{X}_r$). The obtained function is called Spatially Varying Bidirectional Reflectance Distribution Function (SVBRDF). The following sections present the state of the art for the reconstruction of the SVBRDF, Surface Light Field (a further simplification obtained with fixed incident light direction), Reflectance Transformation Imaging and Apparent Diffuse Color, which are the main topics of the algorithms proposed in this thesis.

2.3.1 Spatially Varying BRDF

The Spatially Varying BRDF is a 6D distribution function that describes how each single incident light beam is reflected in the hemisphere of all possible outgoing direction. In the specific for each surface point (u, v) and for a pair of incident (θ_i, ϕ_i) and outgoing (θ_r, ϕ_r) directions, it is equals to the ratio of reflected radiance exiting along the outgoing direction to the irradiance incident on the surface from the incoming direction. Its limited number of input parameters allow the reproduction of opaque surface, where the most relevant effects are the diffuse, specular and back-scattering. Further simplifications of the function can be obtained for homogeneous material (4D function without dependency from the surface position) and for isotropic material (3D function invariant to rotation around to the normal).

In order to be physically plausible a BRDF must respect two important constraints: the Helmholtz reciprocity and the energy conservation. The Helmholtz reciprocity implies that BRDF must remain the same when the incident and outgoing direction are reverted. The energy conversation states that a surface cannot reflect more light than was received.

An alternative parameterization of the BRDF [171] uses the halfway vector between the incoming and the reflected rays and a difference vector, which is just the incident ray in a frame of reference in which the halfway vector is at the north pole. In this way, the axes are aligned with directions of common BRDF phenomena (specular and retro-reflective peaks) and this enables representations that are both intuitive and efficient.

BRDF Representation

There exist different approaches to represent and to model the BRDF (see [53] for a complete overview). The most straightforward representation of the reflection properties is to store the BRDF samples for a discrete set of directions in tabulated form. In order to evaluate the BRDF for a given incident and outgoing direction, the tabulated entries are interpolated. A general approach to reduce the data is to transform it into a lower dimensional function space by factorization. The most compact representation is the use of a reflectance model, where the reflectance for a pair of directions is obtained by evaluating a formula depending on a small number of parameters. The reflectance models can be classified as empirical or analytical

models. In an empirical model, the parameters of the formula are easy to relate to the observation of a material. An analytical model applies basic principles of physics to the surface's microscopic structure to build the reflectance behavior of the surface.

A popular empirical model was developed by Phong in [153]. The model consists of a diffuse term and one specular lobe, but it is neither energy conserving nor reciprocal. The Blinn-Phong model [16] is an improvement of the Phong model to render more realistic reflections, based on the use of the halfway vector between the viewer and the light positions. In both the models, the reflection properties are modeled by three parameters: a diffuse coefficient, a specular coefficient and a specular exponent that determines the dimension of the highlights.

The Ward model [202] is similar to the Phong model except that it uses an exponential function to compute the specular component. This exponential function is parameterized by the average slope of the microscopic surface roughness. The model permits the modeling of anisotropy material by using two different slopes for perpendicular directions on the surface.

Another generalization of the Phong model is the Lafortune model [103]. The model permits the definition of lobes around any axis, where a lobe is an element of the BRDF that both conserves energy and obeys reciprocity. The axes used to define these lobes are the off-specular direction for specular scattering, the normal direction for diffuse scattering and the direction of the light source for backscattering.

A more complex model, the Ashikhimin-Shirley model, was proposed in [8]. In this model, when the incident angle increases, the specular reflection increases while the diffuse reflection is appropriately reduced to maintain energy conservation. To model the specular reflection it uses an approximation of the Fresnel factor, which describes how much light is reflected in function of the incident angle.

Analytical models begin by modeling the surface geometry at microscopic level with statistical methods. They assume that the surface is composed by micro-facets. The model must compute three important measures: how many micro-facets are oriented so that they will reflect light in the view direction; how much of each micro-facet cannot be reached from the incident light and how much cannot be seen by the viewer; how much light the micro-facet reflects.

Popular models are Blinn [16], Cook-Torrance [31], and Oren-Nayar [149]. In Blinn and Cook-Torrance, the micro-facet reflectance is assumed specular. The reflectance is modeled as the product of three factors: the Fresnel factor, the facet distribution model, and a shadowing-masking function. The difference between the two models is the choice of the facet distribution model. In Oren-Nayar, the micro-facet reflectance is assumed diffusive. The model permits the modeling of the backscattering. Usually, Cook-Torrance and Oren-Nayar are putted together for modeling a complete BRDF.

BRDF Measurement

The classic device for measuring a general and anisotropic BRDF is the four-axis gonioreflectometer. This device is a combination of motors used to position a light source and a detector at various locations on the hemisphere above a planar material sample. The detector is typically linked to a spectroradiometer or another optical device that permits recording of dense spectral measurements for each configuration of the light and detector (e.g [204]). An optimized three-axis setup for isotropic material was proposed in [61] (see Figure 2.12) and in [113]. In this case the device requires only three degrees of freedom with a less acquisition time and equipment cost.

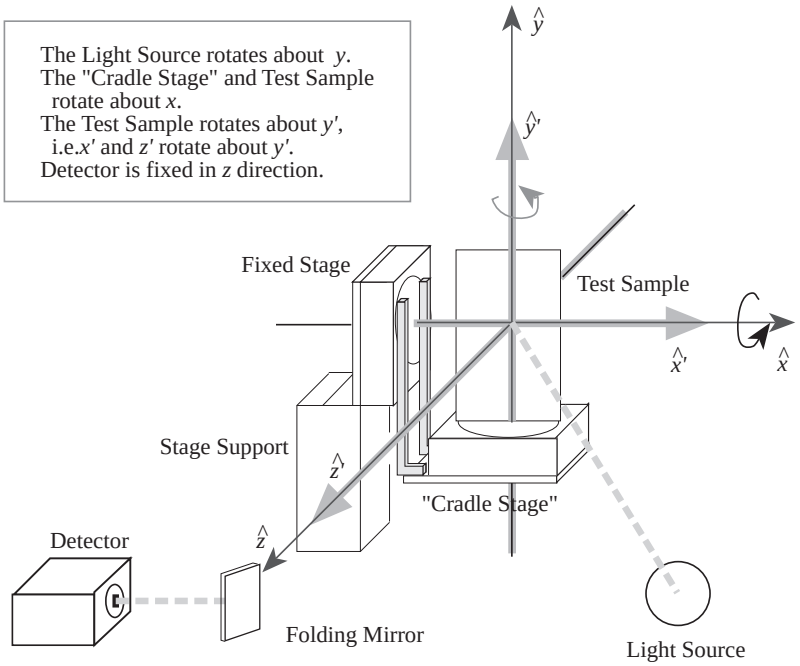


Figure 2.12: *Gonioreflectometer setup by [61]*

BRDF acquisition can be made less costly and time-consuming by using a digital camera, which allows the collection of large number of reflectance samples simultaneously with its sensor.

An early example of image-based system is Ward's measurement setup [202], in which the radiance emitted by a planar sample is reflected from a half-silvered hemisphere and captured by a camera with a fish-eye lens. In this way, a single image provides a dense sampling of the entire hemisphere of output directions and, for a single light source direction, it yields a densely sampled 2D slice of the BRDF. A following solution for obtaining data from multiple directions developed by Bangay et al.[9] uses a parabolic mirror, which allows acquisition of a complete anisotropic SVBRDF without mechanical rotation stages. Ghosh et al. [72] propose a new

setup, without motion and with a digital projector, that measures the BRDF in a basis representation by projecting incident light as a sequence of basis functions from a spherical zone of directions. An alternative configuration is the spatial gonioreflectometer, proposed in [39], that is a gonioreflectometer where sensor is replaced by a digital camera.

Instead of using curved mirrors, it is also possible to use curved geometry to obtain a large number of samples with a single image. Images of curved objects show for each pixel a slightly different normal direction and thus a different viewing and lighting direction in the local tangential coordinate frame of every point. Lu et al. [121] describe an experimental setup for measuring the BRDF of velvet that uses the fabric wrapped around a cylinder. Ngar et al. [143] analyze a solution for measuring the BRDF of anisotropic materials from cylinders where to paste several strips of the material at different orientation relative to the tangent direction. Marschner [128] details a system for measuring BRDFs from spheres and cylinders. Matusik [130] built a system for spherical samples. He captured data for over 100 materials and, from the analysis of the data, he proposed a system to specify and modify the reflectance function. He also proposes a method for more efficient measurement [131].

Other solutions used samples of arbitrary shape, allowing the measurement of the reflectance directly from real object. In this case, an important aspect is the presence of inter-reflections on non-convex shapes, because reflectance values can no longer be directly inferred from intensities observed on the image plane. An extensive overview of the methods for the measurement of SVBRDF of arbitrary shape objects is presented in the next section.

All the methods described above are based on an acquisition with a carefully controlled illumination. An alternative is to recover reflectance information from one or more natural images via inverse rendering. Since inverse rendering problems are ill posed, to solve them it is necessary to make strong assumptions about the materials that are in the scene. One approach is to use a parametric BRDF model. An example is the method proposed by Boivin et al. [17], who employ the anisotropic Ward model and show that when the scene geometry, camera position, and lighting are all known, the BRDF parameters can estimated from one image. In their system, surfaces are manually grouped into regions of homogeneous reflectance, and then BRDF parameters are iteratively updated by comparing rendered results with the input image. In other works, parametric BRDF models have been used to enable the simultaneous recovery of reflectance and shape, or reflectance and illumination. Georghiades [69] presents a method to recover the 3-D shape of surfaces up to the binary convex/concave ambiguity together with its reflectance properties, using single light source images with unknown lighting. Hara et al [82] presents a method for simultaneously estimating the illumination of a scene and the reflectance property of an object from single view images, assuming that the illumination consists of multiple point light sources and the shape of the object is known. By first representing the specular reflection as a mixture of probability distributions on the

unit sphere and then using the Expectation-Maximization algorithm to estimate the mixture parameters, they are able to estimate not only the direction and intensity of the light sources but also the number of light sources and the specular reflectance properties. Ramamoorthi et al. [162] derive an elegant framework for inverse rendering without parametric BRDF models by interpreting the rendering equation as a convolution. This yields an important theoretical tool that, among other things, enables the recovery of reflectance through deconvolution when the scene lighting and surface geometry are known, and when the complete 4D light field is observed.

Image-based SVBRDF Acquisition

The reflection properties of a single material can only partially represent the appearance of a 3D model since many objects have more than a material and may show variation of the reflectance even within a single material. Moreover, adjustments should be made to correct the reflection value from concave portions of the surface that are illuminated both by the calibrated light source and by self-interreflections.

A possible criterion to classify the image-based methods to acquire the SVBRDF proposed until now is the density of the input data in term of angular sampling of the view and light directions (the incident θ_i, ϕ_i and outgoing θ_r, ϕ_r directions).

The multi-view photometric stereo methods recover the reflectance using images captured from different view and light directions. The Bidirectional Texture Functions (BTF) [38] are one of first solutions proposed. A BTF simulates the reflectance effects due to the meso-structure of a complex inhomogeneous material, a structure in-between the macro-scale geometry modeled by triangles and the micro-scale geometry modeled by analytical BRDF. The meso-structure models and plays a very important role in defining and transporting the unique appearance of a material. The BTFs acquisition requires a complex dedicate device working in a highly controlled measurement environment, which allows a dense sampling of the light and view directions to estimate the reflectance of planar samples. This acquisition method is time consuming and data-intensive and, although several solutions have been proposed to reduce the acquisition time and the size of the final reflectance representation (see [138] and [58] for a complete overview), it cannot be extended in an easy way to objects with a complex shape. Examples of compact representations are the non-parametric material model for SVBRDF, based on the concept of shade tree, proposed in [105], or the high-quality general representation proposed by Wu et al. [210] that is, at once, compact, easily editable, and can be efficiently rendered.

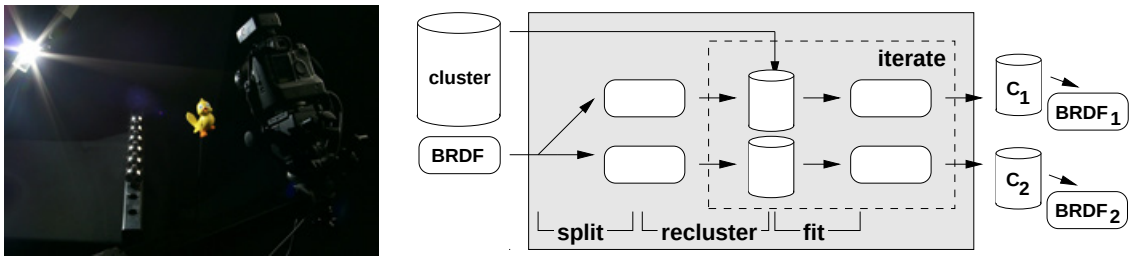
Debevec et al.[44] propose a specialized device setup, the Light Stage, for the relighting of the human face. During the acquisition, the Light Stage illuminates the subject from a dense sampling of directions of incident illumination and recorders it from different angles by stationary video cameras. The new rendering of the subject's face, from the original viewpoints under any incident field of illumination, is computed by linear combinations of the acquired images. Following a geometric model of the face and a skin reflectance model are used to extrapolate the reflectance

observed by the cameras and that would be observed from novel viewpoints. Following extended by Wenger et al.[203] with the possibility to use time multiplexed illumination to speed-up the acquisition, the method requires a careful calibration and a data intensive acquisition. Schwartz et al.[174] use a complex dome setup to acquire the BTF and 3D geometry of Cultural Heritage artwork. The structure of the dome, with 151 cameras and 8 projectors synchronized together by several PCs, imposes strong limitations on the size of the object, usually medium-small object that can be moved from its place.

Several strategies have been proposed to reduce the amount of data needed to acquire a SVBDRF. A class of methods uses polarized light to separate specular and diffuse reflectance. Ma et al.[124] allow the measurement of the diffuse and specular normal maps of an object using four spherical gradient illumination patterns. The estimated normal maps are used in a real-time shading model that allows the reproduction of some subsurface scattering effects. The method is extended in [74] to estimate a spatially varying specular roughness and the anisotropy direction using a set of second order spherical gradient illuminations. Ghosh et al.[73] propose a per-pixel estimation of diffuse albedo, specular albedo, index of refraction, and specular roughness of isotropic BRDFs using few observations of a scene under a single uniform spherical field of circularly polarized illumination. In [75], a new process for multi-view face capture is presented. The key aspect of the method is a new pair of linearly polarized lighting patterns that enable multi-view diffuse-specular separation under a given spherical illumination condition from just two photographs. In general, all these methods require a specialized light dome with an accurate orientation of the camera's polarizing filter.

An alternative approach to reduce the data to acquire is to fit a BRDF model for each surface point using the redundancy of the points that share the same reflectance. Lensch et al. [108] propose an acquisition in a measurement lab with highly controllable lighting conditions (Figure 2.13). They apply a BRDF fitting process that clusters the acquired samples into groups of basis materials that are used in a linear interpolation step to model the final BRDF of each point. The align photo and the 3D model are initially used to create for each visible surface point a data structure called lumitexel, which stores the position, the normal, and the list of the reflectance samples. Then starting from the assumption that the object is spatially homogeneous, a Lafortune BRDF model is computed using a subset of accurately selected lumitexels. Following two clusters are created based on the distance between the measured reflectance values of the lumitexels and their values computed from the estimated BRDF. This spitting process is repeated until the number of clusters is equals to number of materials specified by the user. The final output of this slitting process is a clusterization of all lumitexels in a set of a basic material with the relative Lafortune BRDF. The final BRDF of each lumitexels is obtained as weighted sum of the materials that are representative of the cluster, selected using Principal Function Analysis and some heuristics.

Zickler et al. [217] increase the angular accuracy of a spatially-varying reflectance



(a) (b)

Figure 2.13: *BRDF acquisition by Lensch et al. [108]: (a) acquisition setup; (b) split-recluster-fit process.*

function exploiting the high spatial resolution that images provide to fill the holes between sparsely observed view and lighting directions. The proposed method builds on three principal observations: smooth spatial variation, which allows the use of different pixels for the estimation of the same reflectance function; curved surfaces, which for each image provide 2D slices in the higher dimensional SVBRDF domain; angular compressibility, assuming that a typical BRDF varies slowly over much of its angular domain and varies rapidly only in certain dimensions, such as the half-angle. The final reflectance function is modeled with a non-parametric representation based on radial basis functions. Holroyd et al. [92] present a novel optical setup for measuring the 3D geometry and the spatially-varying surface reflectance of physical objects (Figure 2.14). The basic building block is made by a digital camera and a high frequency, spatially modulated sinusoidal light source that are aligned to share a common focal point and optical axis. Using at least two of these assemblies, they capture a sequence of images of the object from different viewpoints under time-varying sinusoidal illumination originating from different locations. They model the spatially varying reflectance as a low-dimensional subspace spanned by a small set of basis BRDFs, to share the data between different points that belong to the same material.

Haber et al. [80] propose an approach to recover the reflectance of a static scene with known geometry from collection of images taken under unknown and variable illumination. First, they use an all frequency relighting framework, based on a wavelet representation of the visibility and of the current estimation of illumination and scene reflectance, to render the scene using the triple-wavelet product integral. Therefore they employ an iterative optimization to estimate the illumination given the scene reflectance and vice versa.

Another class of solutions is the multi-view methods that use images takes from different viewpoints with a fixed light direction. Nishino et al. [146] propose a method to estimate the reflectance parameters and the illumination distribution from a set of sparse images. They compute three steps: the recovering of a view-independent reflectance map assigning the minimum pixel value projected over each

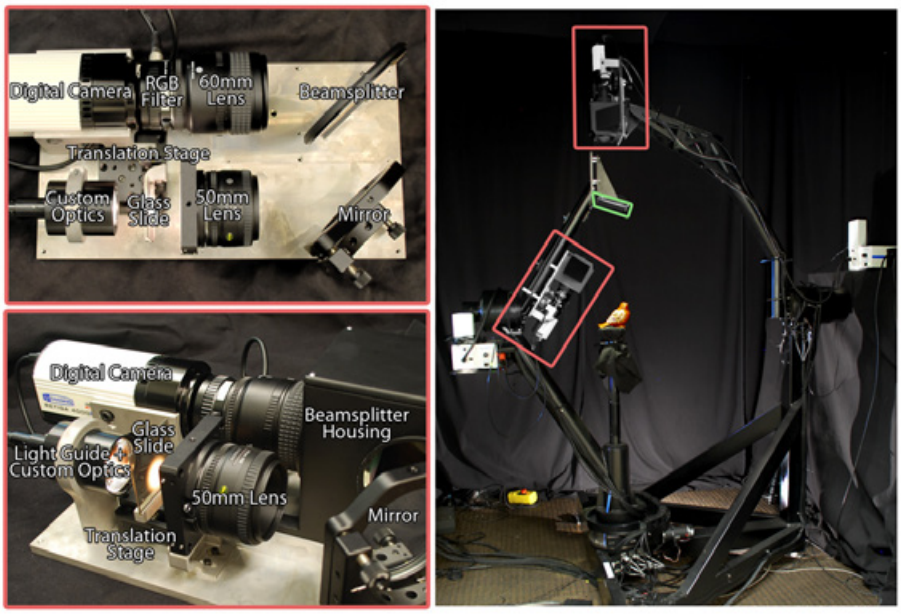


Figure 2.14: *Coaxial optical scanner proposed in [92].*

point of the surface; the initial estimation of the illumination distribution by shooting the residual images along the perfect mirror direction; the iterative refinement of the illumination distribution and the surface reflectance parameters. This method assumes that the specular reflectance is uniform over the entire object. Yu et al. [212] present a method to resolve the usual texture-illumination ambiguity using the view-dependent component of the surface reflectance. The output radiance of the object surface is arranged in a 3D tensor that can be decomposed in three unknown components: the illumination, expressed as combination of spherical harmonics; the light transport, the joint effect of the basis BRDFs and the surface geometry; the texture albedo. All the unknown variables are estimated at the same time by solving a system of bilinear equations obtained from a factorization of the radiance tensor. The reconstruction of the illumination environment map through spherical harmonics decomposition is used in other fields, as the multi-view stereo shape refinement ([209, 208]).

On the opposite side of the multi-view methods, there are the photometric stereo methods where the reflectance function is recovered using different images takes from the same viewpoint but with different light directions. The Polynomial Texture Map [127] is an example of image-based compact representation of photometric stereo data, where for each pixel the reflectance behavior is modeled by a biquadratic polynomial without the need of a 3D geometry. Starting from the observation that many objects can be decomposed into a small number of materials, Goldman et al. [77] present a method for the recovering of shape and SVBRDF. They set an iterative system initialized using the Lambertian photometric stereo data. The

first step estimates the BRDF parameters of the basis materials. The second step estimates the per-texel normals and the material weights. The third step computes the 3D surface using the normals to solve a Poisson equation. A similar approach is used in [7] with a more general model based on a novel bi-variate approximations of isotropic BRDF.

A number of solutions have been proposed for the photometric stereo acquisition of flat target moving a linear light source. Gardner et al. [66] recover both the geometry and the reflectance by combining a laser strip scanning with two scans with the light source in a diagonal orientation. From the camera and light positions at each image, an isotropic Ward model is fit to each pixel. In [200], the system is extended to allow the estimation of anisotropic reflection by replacing the linear light source with a LED array and by merging the similar reflectance data of different pixels. Ren et al. [164] propose a portable setup device with an additional element: a BRDF chart, a small card with a set of known BRDF reference tiles (Figure 2.15). The light responses from the char tiles as well as from the points on the target are acquired and following matched with dynamic time warping techniques to recover the target's appearance. The algorithm works with LDR image sequences, without knowing the light and view directions.

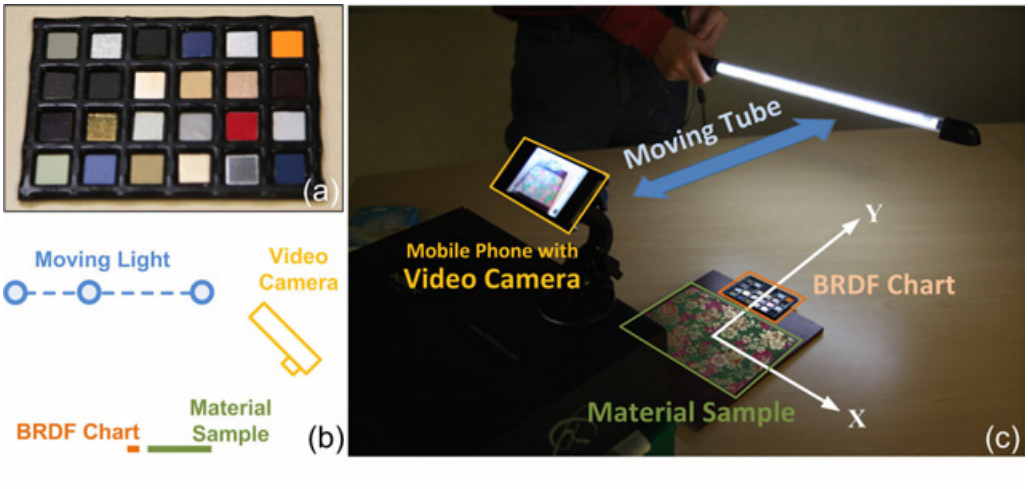


Figure 2.15: *Device setup with BRDF chart proposed in [164].*

The idea of using some reference objects of homogeneous reflectance and known geometry to reconstruct the appearance of a target object without any knowledge about the view and the light directions was introduced by Hertzmann et al. in [89]. Treuille et al. [190] extend the method to reconstruct the whole object from multi-view data. The main drawback of these methods is the need to have reference BRDFs similar to the target's reflectance. To relax this requirement some methods perform a two-step acquisition on the object. To obtain surface reflectance of large outdoor scenes, Debevec et al. [45] measure a set of representative BRDFs from small regions of the scene using controlled lighting, as well as images of the entire

scene under natural lighting. At each scene point, the Lambertian color is recovered and its BRDF is modeled as a linear combination of two representative BRDFs whose diffuse colors are the most similar to the point. This approach fails in general when surface points have similar diffuse colors but different specular reflectance. Dong et al. [51] present another two-pass method. In the first phase, they acquire some representative BRDF samples, sparsely distributed over the surface with a dense angular sampling, using a fast single-point BRDF measurement device. In the last phase, they acquire a sparse angular sampling of the appearance of each target point with a set of images to constrain the final reconstruction of their reflectance as a linear combination of the nearest representative BRDFs.

2.3.2 Surface Light Field

All view-dependent surface appearance representation can be derived from the plenoptic function which was introduced by Adelson [2]. It represents the radiance at every point in a scene (x, y, z) in every direction (δ, ϕ) depending on wavelength λ and time t. For static scenes, the time may be neglected, and the dependency on the wavelength is frequently reduced to three RGB samples.

Two equivalent realizations of the plenoptic function were proposed in form of the Light Field [112] and Lumigraph [78]. Both are five-dimensional functions that depend by the surface location and the viewing direction. However, the 5D representation may be reduced by one dimension in free space since radiance does not vary along one ray until it hits an obstacle. Hence, a 4D parameterization can be given to all possible rays by restricting all point to lie on a convex hull of the considered object.

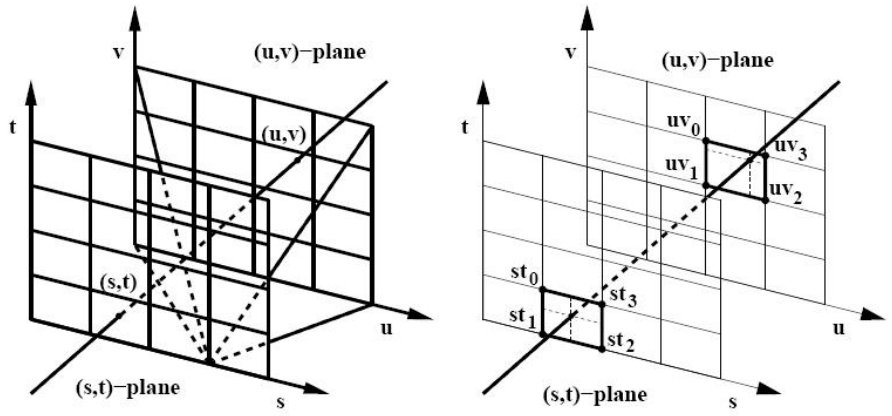


Figure 2.16: *Two plane parameterization of the Light Field.*

In [112], the function is given by a two plane parameterization, where all possible rays are parameterized by their intersection points with two parallel planes with plane coordinates (s, t) and (u, v) (Figure 2.16 and 2.17). For each ray, the light

field stores the radiance traveling along the ray. The (s,t)-plane is the camera plane in which all the camera are placed. The (u,v)-plane is the focal plane. Both planes are typically discretized into a regular grid, storing radiance values only for the rays passing through the grid points. An arbitrary ray intersecting one cell on the (s,t)-plane and another cell on the (u,v)-plane is surrounded by 16 rays (Figure 2.16). The desired radiance value is obtained by quadri-linear interpolation of these 16 samples. For high-resolution light fields this interpolation technique works fine. However, if the resolution is decreased, only points in the (u,v)-plane will appear sharp, while blurring will increase with respect to the object's distance from the image plane. To solve this problem Gortler et al. [78] augment the light field by some geometric information. A coarse polygon mesh approximating the object is used to determine the depth along the sampling rays. Based on this depth information, the weights for the interpolation are corrected. In this way, sharp contours of objects even for low resolutions are obtained.

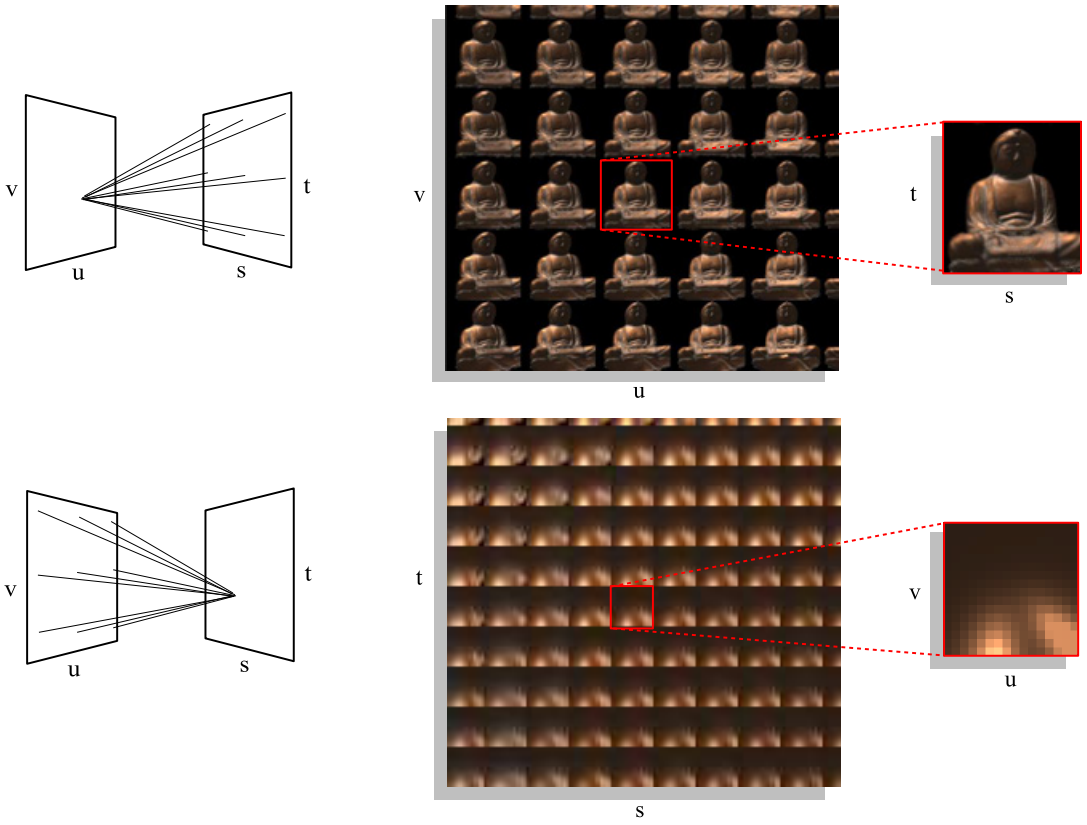


Figure 2.17: *Two different visualizations of a Light Field from [112]*

The Surface Light Field was introduced in [134] to sample Light Field on a parametric surface directly. There, the light field is closely coupled to the geometry of the reflecting object. The Surface Light Field is a function that assigns a color to each ray originating on a surface and it is well suited to generate virtual images

of objects under complex lighting conditions. The parameters u and v are chosen to match the surface parameters, thus every sample in the light field is associated with a location on the surface. The remaining s and t parameterize the hemisphere of directions above the surface point (u, v). The use of the geometry of the object allows overcoming some of the drawbacks of a generic Light Field representation: the camera is restricted to certain regions of space; the finite angular resolution leads to depth of field effects; objects become blurry in proportion to their distance from the image plane. However, some rendering artifacts appears if the directional resolution is low, because the resolution in s and t directly affects the quality of the specular highlights rendering. Furthermore, light fields can only capture the view-dependent effects and always show the object in the same lighting environment where it was during the acquisition. It is not possible to adapt the incident lighting to new virtual environments.

One of the most important problem in the Light Field and Surface Light Field acquisition and estimation is the not uniform sampling of the view direction. In the literature, this problem has been approached in two different ways: hardware-based and software solutions.

In the hardware-based approach, the proposed methods use different type of special hardware devices that allow the acquisition of a uniform sampling of the view direction. Levoy et al. [112] use a video camera mounted on a computer-controlled gantry with a planar camera motion in front of the object. Since the success of the method depends on having a high sample rate, they use a two-step compression system, based on vector quantization and entropy coding, in order to make creation, transmission, and display of light fields practical. Shum et al. [178] propose to constrain the camera motion to planar concentric circles in order to create concentric mosaics. They obtain much smaller file size because only a 3D plenoptic function is constructed. Other solutions are based on the use of camera arrays (Figure 2.18): the 6-camera system for on-the-fly processing of generalized Lumigraphs, which is a two-plane parameterized Lumigraph with per-pixel depth information proposed in [173]; the system of 16 cameras for real-time rendering, proposed in [139]; the Distributed Light Field Camera [211] that is able to render live dynamic Light Fields from an 8×8 array of commodity webcams; the Self-Reconfigurable Camera Array [213] with 48 commodity Ethernet cameras with electronic horizontal translation and pan controls to improve view interpolation results on-the-fly; the array of 100 custom synchronized video cameras [205] to allow the capture of a video light field, to which it is possible to apply spatio-temporal view interpolation. Similar methods use different additional devices: microlens array [142, 70], to put between the sensor and the main lens of the camera to sample the 4D Light Field on a single photographic exposure; additional optics [68], with systems of lenses and prisms used as an external attachment to a conventional camera that sacrifice angular resolution for a higher spatial resolution; attenuating mask [196] [195], to put into the optical path of a lens to attenuate the light rays inside the camera instead of bending them.

In the software approaches, the main idea is to use algorithmic solutions to fill

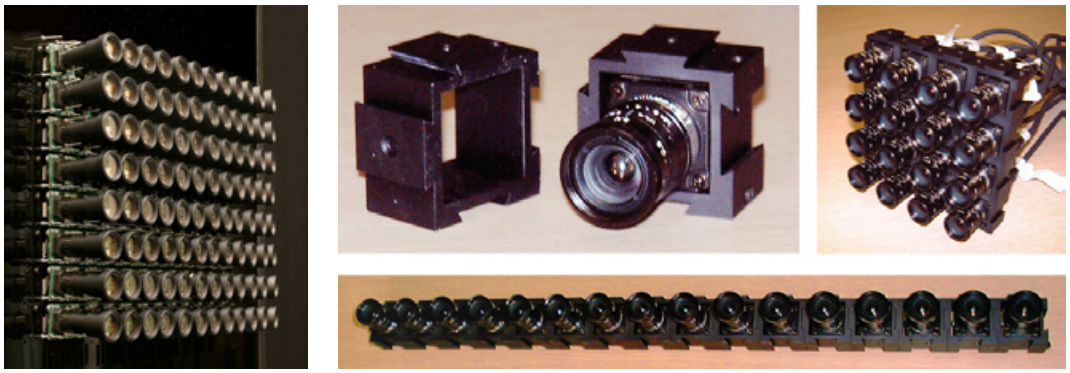


Figure 2.18: *Camera array examples from [205] and [139].*

the gaps and the holes coming out from a sparse or not spatially uniform directional acquisition. Gortler et al.[78] use a hand-held camera with a specially-designed stage for camera pose estimation. They rearrange the acquired samples in a two plane parameterization and use a pull-push algorithm to fill the gaps on each plane due to a not uniform distribution of the acquired view directions. Heigl et al.[87] start from an image sequence taken by a hand-held camera and relaxed the restrictions imposed by the regular Light Field structure. They render new views directly from the calibrated images with the use of a local depth map, mapping directly the original images onto one or more planes viewed by a virtual camera. Isaksen et al.[96] propose a new parameterization, based on a general mathematical formulation for a planar data camera array, which allows representation of moderately sampled Light Fields with wide variations in depth, without requiring geometry. The analytical formulation of the minimum sampling rate for Light Field rendering is derived in [26] using the sampling theorem and the Fourier spectral analysis of the Light Field signals. Buehler et al. [22] propose a generalization of two image-based rendering algorithms: Light Field rendering and View-Dependent Texture mapping [46]. In particular, they allow a lumigraph-style rendering from a set of input cameras in arbitrary configurations. In the case of regular and planar input camera positions, the algorithm reduces to a typical Lumigraph approach. When presented with fewer cameras and good approximate geometry, the algorithm behaves like View-Dependent Texture mapping. Chen et al. [27] propose the approximation of the Light Field data by partitioning it over the mesh vertices and factorizing each part into a small set of lower-dimensional functions. They rearrange the input data into a matrix and then computed an approximated factorization of the matrix with two different methods: Principal Component Analysis and Non-negative Matrix Factorization. In order to obtain a more precise factorization, they apply a resampling of the data by Delaunay triangulation of the original views in the xy plane of the local reference frame of the vertex. This resampling step requires a good distribution of the original view in the local reference frame, a condition that a irregular acquisition

does not guarantee. Davis et al. [41] presented a system for interactively acquiring and rendering a Light Field using a hand-held commodity camera. They compute and provide a coverage map to the user in real-time to show the views captured so far and to help the achieving of a dense coverage. A visual feedback mechanism for capture guidance is used also by Jachnik et al. [97]. They propose a real-time system for the capture of the Surface Light Field of planar surface based on the splitting of the diffuse and specular components. The method assumes an uniform specular behavior on the whole surface, modeled as a Phong lobe, and cannot be easily extended to objects with a more complex reflectance.

Wood et al. [207] propose three different approaches for the Surface Light Field estimation. The first step in the estimation process is the construction of a useful intermediate representation, consisting of a data lumisphere for each grid point in the Surface Light Field (Figure 2.19). A data lumisphere is a set of samples, each consisting of a color and a direction corresponding to an observation of a grid point. The first approach is the construction of a piecewise-linear lumisphere from each lumisphere data independently at each surface point. It uses a regularized error function that is the sum of a least-square term and of a fairing term, based on a discrete approximation of a thin-plate energy. This approach shows its limitations when the holes in the acquired view directions became too big. To generate a more compact Surface Light Field, they propose two other methods, which represent each lumisphere as a weighted sum of a small number of prototype lumispheres: the function quantization, analogous to vector quantization; the principal function analysis, analogous to principal component analysis. To improve the compression performance they apply a median removal algorithm to store the median texture map and to encode the residual separately. These approaches cannot be generalized very well for not uniform directional sampling because they require some prototype lumispheres with a good and uniform directional sampling, that are difficult to obtain with simple camera paths.

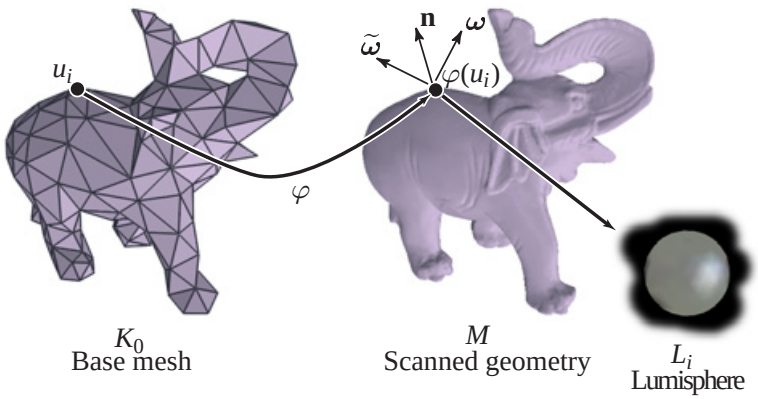


Figure 2.19: *Representation of the surface light field in [207].*

2.3.3 Reflectance Transformation Imaging

Reflectance Transformation Imaging (RTI) is a computational photography method that, stating from a set of images taken from a single view under varying lighting conditions, encodes the surface shape and color to enable the interactive re-lighting of the subject from any direction. RTI encodes this redundant data, the same scene sampled under many different lighting conditions, in a compact way, using view-dependent per-pixel reflectance functions.

Polynomial Texture Map (PTM) was the first proposed RTI image format [127]. This image encodes the per-pixel reflectance function using a biquadratic polynomial. Two different image formats are proposed: LRGB PTM that stores a pseudo-albedo RGB color and the six coefficients of the polynomial for the luminance channel (9 byte per pixel); RGB PTM that store the coefficients of the polynomial for each color channel independently (18 byte per pixel). Already in its original work, the PTMs are presented as a powerful tool to improve the study of ancient writings and inscriptions. Authors propose two contrast enhancement methods to improve the readability of the ancient inscription based on the mathematical manipulation of the biquadratic polynomial. A new encoding for RTI data was proposed by Gunawardane et al. [79] with the Hemispherical Harmonics Map. This new format uses a linear combination of the first nine hemispherical harmonics to model the reflectance function (nine coefficients for each color channel then 27 byte per pixel), enabling a higher rendering quality due to the better sampling properties of the new per-pixel function.

In the last 10 year, RTIs have been employed in several Cultural Heritage projects. They were used in Paleontology, to provide noticeable improvement in imaging of low color contrast, like high relief fossils [81]. The application of PTM method on ancient stone tools revealed fine details of conchoidal knapping fractures, use scarring and stone grain [136]. A joint work done by National Gallery and Tate Gallery of London showed that PTMs provided additional information about the surface textures of oil paintings [150]. The application of PTMs and scanning techniques on a large numismatic collection permitted the creation of a more complete documentation than the traditional photographic methods and the communication of this information with ease through digital media [137]. PTMs were used to study the Antikythera mechanism [64], an ancient mechanical computer designed to calculate astronomical positions. Here, the analysis of the different fragments using PTMs increased the readability of the inscriptions, allowing a more complete understanding of the mechanism operation.

Improved methods have been proposed for the acquisition of RTI images in addition to the classical light dome [127] [206]. Dellepiane et al. [48] present a method to acquire the RTI of large object, using a single moveable light and an acquisition plan without the employ of a light dome. An automatic method for RTI generation is proposed in [10]. It is based on the tracking of the highlight position produced by the light source on a glossy sphere. In this way the user is free to

move a single light source around the object and, after the acquisition, he can use an automatic tool to estimate the encoded light direction of each photo.

Shading Enhancement

Recently, improving the readability of computer-generated images has gained significant interest and many different approaches have been proposed to provide better and more flexible visual inspection capabilities. On the other hand, little attention has been devoted to investigate these issues in the RTI domain.

Especially, the use of different light directions on the different areas of the image is not explored. This approach has been traditionally used by illustration artists to depict more clearly the shape and the important features of a subject. Various authors exploited this idea to build interactive, paint-like systems that combine, in a user driven way, different portions of multiple lighted images [5] [4]. The main idea is to create a better image by combining portions of images from a collection in which the viewpoint and the scene remain fixed, but the lighting varies from image to image. Fattal et al. [55] extend this initial idea using the shading at multiple scales across the input images to generate the enhanced results in an automatic way. For each input image, they compute a multiscale decomposition by bilateral filtering and construct the enhanced output image by combining detail information from all of the input images at each scale of the decomposition. A Non-PhotoRealistic (NPR) shading model is proposed by Rusinkiewicz et al. [170], inspired by techniques for cartographic terrain relief. The effective light position is dynamically adjusted for different areas of the surface, using the local curvature of the surface along with smoothed surface normals as key elements to drive the choice of the new light position. Toler-Franklin et al. [187] introduce RGBN images. An RGBN image stores an albedo color and a surface normal at each pixel. This imaging method enables a user to choose from a variety of filters that are then applied to RGBN data to generate NPR renderings, such as in [170], to enhance the visibility of selected surface features. Vergne et al. [198] enhance the shape depiction of 3D objects by locally warping the environment lighting around the main surface features. The method extracts the salient surface features with an analysis of the normals field in a view-dependent fashion and then it compresses or stretches locally the illumination directions with a warping function that deforms the incoming light direction around these features.

2.3.4 Apparent Color

Starting from a set of photos calibrated with respect to the 3D model, the apparent color value is mapped on the model by applying inverse projection to transfer the color data from the images to the 3D surface. There are numerous difficulties in selecting the correct color to be applied when multiple candidates are present among different images. The most important one are dealing with the disconti-

nuities, caused by color differences between photos that cover adjacent areas, and reducing the illumination-related artifacts, due to shadows and highlights. Then, the two main issues in this process are how to save the color on the 3D surface and how to select the most correct color.

For the storing, there are two possible approaches: texture-based and vertex-based encoding. The texture-based approach requires a mesh parameterization to produce a new texture map, either by joining subregions of the input images or by resampling [13][19][141][107][24][18]. Unfortunately, the management of very dense geometric and photographic sampling is very complicated. The texture-based approach is ideal with moderate resolution meshes (al most 1M faces) and moderate pixel datasets (1-5 Megapixel). Moreover, multiresolution encoding is usually a need for huge meshes, and the adoption of multiresolution approach for texture-based representation of the color [18] implies the need of a multi-resolution texture atlas, increasing the redundancy and the space occupancy.

The vertex-based approach requires the saving of a color value for each vertex of the mesh, while the color inside each triangle is obtained by barycentric interpolation of the three colors of the triangle's vertices. This solution is better for all cases where both a high resolution geometry and photographic data are available. An advantage of per-vertex encoding is the space efficiency, but it is less accurate when the color has a level of detail greater than the geometric detail.

The other issue is the selection of the most correct color. A standard approach is the computation of a texture by assembling subparts of the original input images, posing the problem as an image stitching problem. Additionally, some corrections can be applied to deal with inconsistencies in the borders between different images. Rocchini et al. [166] propose an approach to produce a smooth join between different images that map on adjacent sections of the surface based on four steps: vertex-to-image binding, to detect for each mesh vertex the subset of images that project a valid color and to set the most orthogonal one as target image; patch growing, to change the vertex-to-image links in order to obtain larger contiguous sections of mesh that map to the same target image, for an optimal texture patching; patch boundary smoothing, that applies a local registration to all vertices of the faces on the border between different target images and resamples a new triangular texture patch for each of these faces, computing a weighted composition of the corresponding triangular sections in the associated target images; texture-patches packing into a single texture map. Camera orthogonality is used also in Lensch et al. [107] to choose in which photo each part of the 3D model is mapped. The images are then fully blended, using the entire local redundant area. A further extension is proposed by Callieri et al. [24]. The mesh is covered using the most orthogonal image for each mesh portion and redundancy is used to correct color discontinuities in the boundary between images. The most important novelty is the use of a correction-map that quantifies how many corresponding colors differ on the triangles on the border. This correction-map is used to propagate the color correction factor on the whole texture.

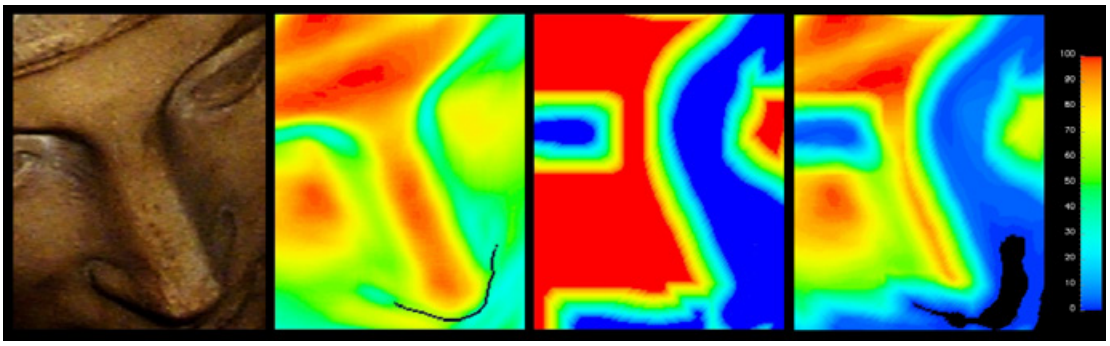


Figure 2.20: *Weight map computation proposed in [14]. From left to right: color image; weight map based on the ratio of projected to true area for each pixel; weight map based on the photometric calculations; final weight image.*

An alternative approach is the use of a per-pixel blending function. Debevec et al. [46] introduce the view-dependent texture mapping. The final color is the weighted average of the color projected by each views and the weights are inversely proportional to the magnitude of the angle between the view vector of the photos and the view vector of the virtual view. An extension to allow rendering of complex geometric meshes at high frame rates without blurring or skinning artifacts is proposed in [159]. Bernardini et al. [14] compute for each image a weight that is the product of two maps (Figure 2.20): the ratio between the projected area and the real area of each pixel, computed as the cosine of the angle between the surface normal and the camera view direction divided by the square of the distance from the camera; a photometric calibration map with a weight proportional to the quality of the reconstruction of the surface normal using only photometric data.

Baumberg et al. [11] compute a weight for each triangle, as the ratio between the projected and the real area of the triangle, project the triangle in camera space and then they apply a 2D Gaussian smoothing to obtained seamless weight map. Callieri et al. [23] present a flexible weighting system where different metrics are used to define a per-pixel weight. The main metrics include the angle between the view direction and the surface normal, the distance from the camera, the distance from the discontinuities in the depth map, the sharpness of the pixel, but the system could be easily extended in order to accommodate additional metrics (Figure 2.21).

The drawbacks of the blending methods are the ghosting and blurring artifacts due to small error in the alignment of the images to the 3D model. To solve this problem, Lempitsky et al. [106] start by back-projecting original views onto the 3D surface. Then a texture mosaic is created from these back-projections, whereas the quality of the mosaic is maximized within a process of Markov Random Field (MRF) energy optimization. Finally, the residual seams between the mosaic components are removed with a similar procedure to the gradient-domain stitching techniques proposed for image editing. Ran et al. [65] extend this method with two improvements: in the MRF optimization they search not only over the images sources,

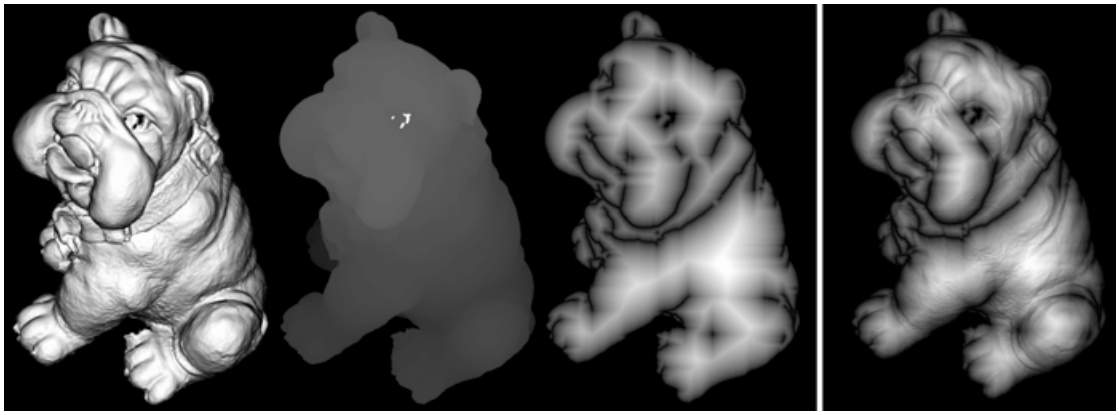


Figure 2.21: *Weighting masks proposed in [23]. From left to right: Angle Mask, Depth Mask, Border Mask, Final Mask.*

but also over a set of local image transformation that compensate for geometric misalignment, typically local image translation; they apply a Poisson blending in texture space to eliminate the residual lighting variations.

In order to reduce the blurring and ghosting artifacts, the Floating Textures system [54] uses optical flow to obtain warp data between images, and then combining the warp fields linearly in the space of the current viewpoint. In this way it is possible to work in real time, but the linear and view-dependent combination of the warp fields prevents the creation of a globally valid texture for the whole object. Dellepiane et al. [49] propose an alternative method, based on the computation of the optical flow between overlapping images, to correct the local misalignment (Figure 2.22). The basic idea of the algorithm is to warp locally the input images in order to minimize small-scale misalignment of high frequency color features, thus obtaining a sharper color mapping. The proposed solution has two components: a method to determine the warps between pair of overlapping images using optical

Figure 2.22: *Results obtained with flow-based local optimization from [49].*

flow; a strategy to combine the resulting image warps to obtain a coherent warping to be used in color mapping.

In order to remove the illumination-related artifacts, a new solution based on flash-based photography is proposed by Dellepiane at al. [47]. The method is based on two steps: a simple procedure to estimate the flash position with respect to the camera lenses and a color correction space, where a color correction matrix is associated to each point in the camera field of view; an automatic method that, using the info of the previous step, is able to improve the integration of the color samples by not taking into account the specular highlights and the shadows contained in the acquired images. The solution works only if the flash-light is predominant in the scene.

Chapter 3

Geometry-Aware Video Registration

The estimation of the motion of the camera from a video has been extensively studied in the Camera Tracking field. The trivial use of the output data obtained by these solutions to solve the Video-to-Geometry registration problem leads to an insufficient accuracy for the reflectance computation that can be solved by integrating the Image-to-Geometry solutions with the two main characteristics of a video: the high frame-to-frame temporal coherency and the data redundancy.
This chapter introduces a new solution for the accurate and efficient alignment of a video sequence of a real object over its digital representation (a dense triangular mesh). The proposed solution uses two different Image-to-Geometry registration approaches: feature-based registration by KLT video tracking; statistic-based registration by maximizing the Mutual Information between the gradient of the frame and the gradient of the rendering of the 3D model with some illumination related properties, such as the surface normals and the ambient occlusion. This chapter also presents a comparative study between the proposed registration by Mutual Information using gradient maps and the original algorithm by Corsini et al. [34] to evaluate the performance improvements.

3.1 Video-to-Geometry Registration

The camera tracking problem has been extensively studied in the last few years and several interesting and robust solutions have been proposed. The general strategy is to identify and track the most salient 2D features of the video and to use these features and their trajectories to recover the motion of the camera and some sparse 3D information about the scene.

Due to the main aim of these techniques, which is to provide a way to render additional elements inside a real-world video, the camera motion and scene information recovered by these approaches are correct up to a scale factor that is difficult

to evaluate due to the characteristics of the scene and of the camera motion. Additionally, in most cases, this scale is non-linear and changes in time and even across the scene. While, using this type of data, it is possible to render a 3D model as an additional component of the scene, the alignment of the reconstructed camera motion to a dense 3D model, to be able to project 2D data to the 3D scene, is very complex.

There exist two possible straightforward approaches to use the camera tracking data to solve the Video-to-Geometry registration problem. Starting from the manual alignment of the first frame over the 3D model, the first solution can use the frame-to-frame coherence to compute a set of 2D correspondences between pair of consecutive frames with a KLT tracker. These 2D correspondences can then be used by an automatic procedure that, starting from the registration of the previous frame, extracts a set of 2D-3D matches and recovers the camera parameters of all frames in an incremental way, using a non-linear least square optimization. Unfortunately, this solution has a problem due to the drift errors introduced by the sequential KLT processing. The drift problem is due to several causes: image noise, geometric distortion, illumination changes, occlusions, fast camera movements, 3D features that leave the camera field of view and later reappear in the sequence (often the same 3D point can be found and tracked again later in the sequence, but a new track with a new 3D projection is made). This kind of drift is particularly challenging in long sequences, where the camera visits the same part of the scene multiple times, and it leads to a misalignment of the 3D model with respect to the video frames.

The second solution can try to align the sparse 3D point cloud computed by the camera tracking with the dense 3D model and to transfer the transformation to the camera position and orientation of each frame. Even if several solutions have been proposed to solve this 3D registration problem, the computation of good alignment is not guaranteed for several reasons, like noisy point cloud, not uniform scale factor between the 3D models, no metric reconstruction. Then the final alignment of the video on the 3D model is not very accurate and requires further refinement steps to improve it (like in [33]), that do not take advantage from the redundancy and frame-to-frame coherence of the video.

On the other side, the alignment (registration) of a 2D image on a 3D model is a very well know issue in the Computer Graphics field. Different solutions, both semi-automatic and completely automatic, have been proposed in the late years, which are able to align images to dense geometries coming, for example, from 3D scanning. However, despite the availability of such methods, the trivial idea of applying the semi-automatic or even the more automatic methods for 2D-to-3D registration to each frame of the video would result in a high computation time.

Given the amount of works in the 3D Computer Graphics field that make a profitable use of 3D registered images to enrich digital models, being able to exploit the advantages of video sequences (frame-to-frame coherence, redundancy of data) could be a great help in different applications. If an accurate registration of the video on the 3D model is obtained, the bi-directional data transfer could be used

for a number of interesting applications (color transfer, estimation of reflectance properties, recording of appearance-varying scenes). Up to now, no solutions have been proposed to align accurately a video sequence over a mesh using the redundancy and the high frame-to-frame coherence of the video.

This chapter presents a method to align, efficiently and accurately, a video sequence to a dense 3D geometry, combining the speed and flexibility of the feature-based tracking and the high precision and geometrical consistency of the image registration approaches. The proposed method combines the KLT tracking with a state of the art image registration technique based on Mutual Information [34], a statistical measure of the information shared by the image and the rendering of the model. These two approaches can be considered orthogonal, since they deal with different information extracted from the data (feature vs. statistical analysis). Both the approaches are needed because the MI corrects the drifting effect that the KLT tracking produces over long sequences, due to the incremental tracking and the camera motion, while KLT tracking speeds up the registration, allowing a fast registration of short sequences with simple camera movements, and controls the convergence of MI towards good camera parameters.

3.2 The Geometry-Aware Registration Algorithm

The algorithm computes the camera parameters for each frame taking in input a dense triangular mesh of the object and a video sequence acquired with a constant zoom factor. The algorithm assumes a perspective camera model defined by intrinsic and extrinsic parameters. The intrinsic parameters are assumed constant for the whole video sequence and they are estimated only for the first frame in a preprocessing step. More specifically, the skew factor is assumed equal to zero, the principal point is set as the center of the image, the horizontal and vertical scale factors are computed from the image resolution and the CCD dimensions, while the focal length is assumed constant and it is estimated only for the first frame and propagated to the whole sequence. In order to increase the robustness of the tracking of the 2D features, the lens radial distortion is estimated only once, using a single frame of a black and white checkerboard to automatically extract the position of the corners to give in input to the camera calibration method defined in [192] in the case of coplanar points. The extrinsic parameters define the rotation matrix, parameterized by the Euler angles $(\theta_x, \theta_y, \theta_z)$, and the translation vector (t_x, t_y, t_z) that are needed to transform the camera coordinate system into the world coordinate system.

The proposed algorithm is composed by two integrated tasks, the feature-based registration and the registration by MI, preceded by a preprocessing step to compute the camera parameters of the first frame and to extract the 2D features tracks from the video, in order to have a set of correspondences between 2D features for each pair of consecutive frames. The scheme overview of the algorithm and the integration of the two registration tasks are shown in Figure 3.1.

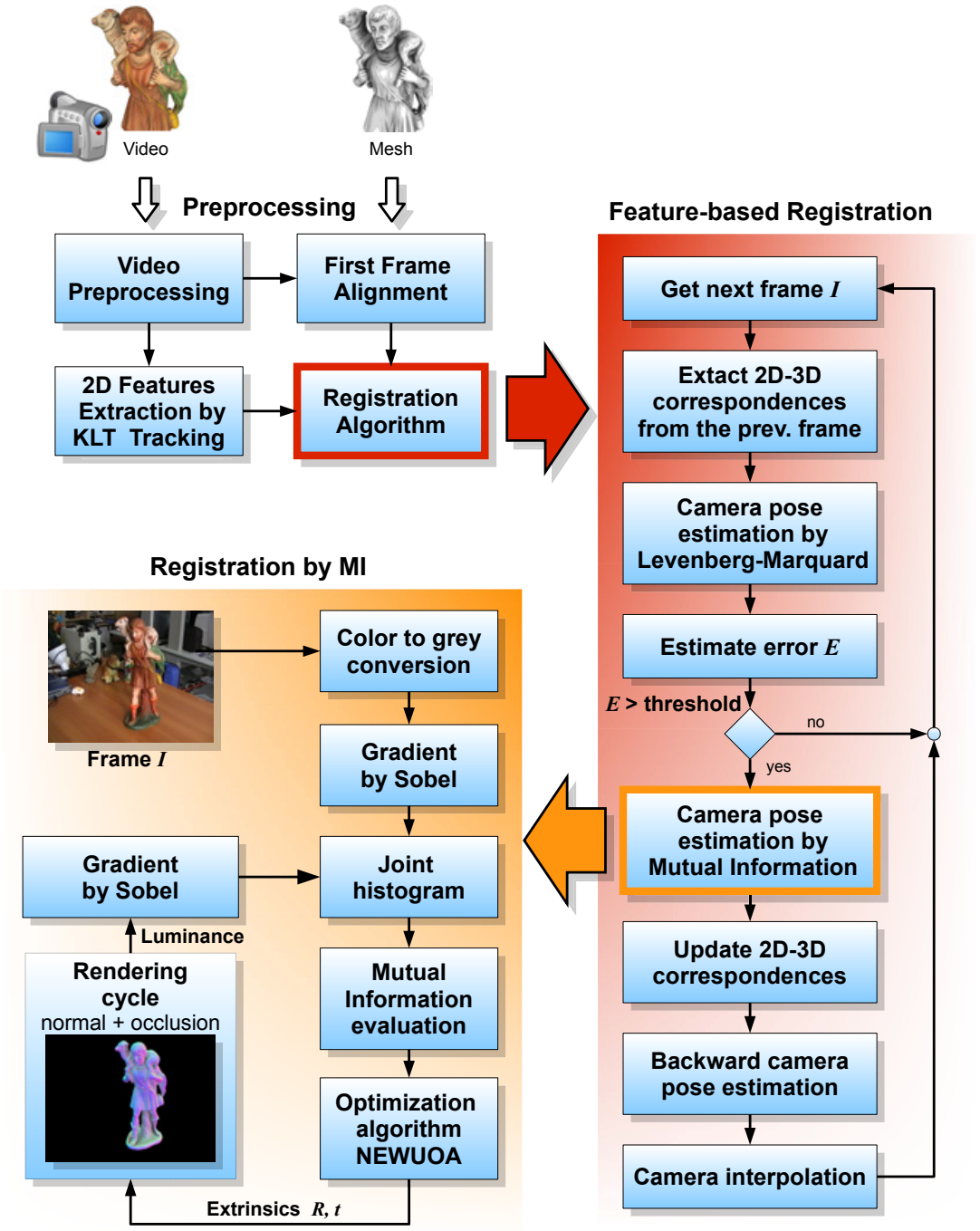


Figure 3.1: *Video-to-geometry registration: algorithm overview.*

3.2.1 Preprocessing

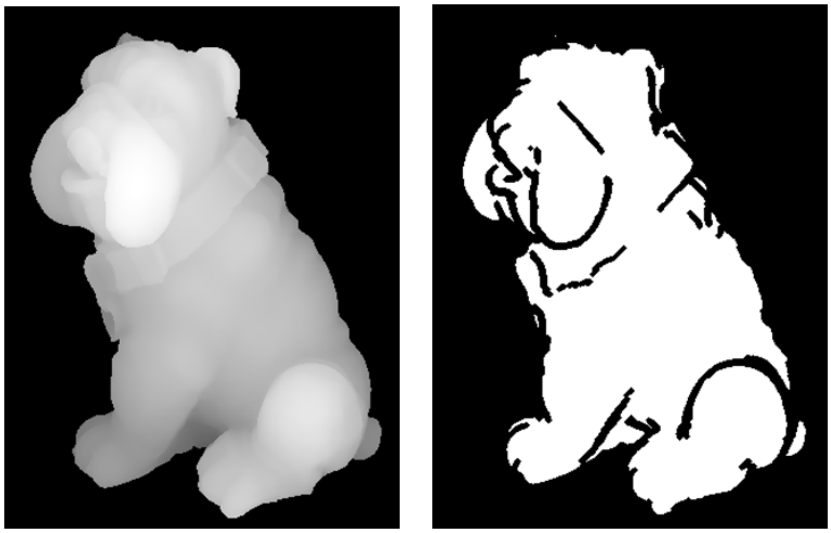
The task of the preprocessing step is to prepare and produce the data needed for the registration procedure. The outputs of this step are the camera parameters of the first frame and the 2D features tracks extracted from the video. The first phase is the application of some video filters to improve the quality of the video data, in order to allow a more robust 2D features tracking. The video is deinterlaced (if necessary), and noise is removed by bilateral filtering. Then, the radial distortion introduced by the camera lens is eliminated from all frames.

The processed frames are used by the following phases to compute the camera parameters of the first frame and the 2D features tracks. The first step is the alignment of the first frame over the 3D model by manual selection of a set of 2D-3D correspondences to use in the Tsai's calibration method, using the tool described in [63], followed by a further alignment refinement with the Mutual Information framework described in [34]. In this way, the focal length and the extrinsic parameters of the first camera are computed. The second step is the extraction and saving of the 2D feature tracks of the video by using the Voodoo Camera Tracker tool [184]. This tool uses a KLT tracker to detect and track the most salient 2D features and it applies a RANSAC approach to make more robust the estimation of the fundamental matrix, useful to discard the outliers. Even if the tool computes also the camera perspective matrices and a sparse 3D scene reconstruction for short sequences (at most 400 frames), the algorithm does not use this information because they are estimated up to a scale factor not easy to find.

3.2.2 Registration algorithm

The registration algorithm works in an incremental manner: to align the frame F_i, it uses the data from the registration of the frame F_{i-1}. From the camera parameters of the previous frame and the 2D features tracking information, it extracts a set S of 2D-3D correspondences. This set S is used to setup a not linear least square problem for the computation of the camera pose with the Levenberg-Marquardt algorithm [118]. For the extraction of the set S a validity mask is computed from the depth map of the frame F_{i-1}. This mask is prepared by applying a Sobel filter to the depth map and then computing a distance field from the border pixels. A pixel is set as a border if its gradient magnitude value is above the 95% percentile of the whole frame. All the pixels that are at most two pixels far from a border are marked as invalid (Figure 3.2). This mask allows discarding all the 2D features of the frame F_i with a corresponding 2D point in the previous frame F_{i-1} that does not belong to the object or that lies near to depth discontinuities, in order to discard wrong 2D-3D correspondences due to small registration error. Then for each valid 2D feature, the procedure computes the 3D point by backward projection onto the 3D model of the corresponding 2D features in the frame F_{i-1}.

To estimate the quality of the registration, given the set S of 2D-3D correspon-



(a) **(b)**

Figure 3.2: *Validity mask for 2D features selection: (a) depth map; (b) validity mask.*

dences m, M and the camera projection matrix P, the following registration error E is computed:

$$E = \frac{1}{|S|} \sum_{\langle m,M \rangle} d(M, P^{-1}m) \tag{3.1}$$

where the function d computes the geometric distance between the 3D point assigned to the 2D feature by the previous frame and the 3D point computed by backward projection of the 2D feature with the camera P onto the 3D model. The error E is computed as the average distance of 3D points keeping constant the 2D features positions. The averaging helps to have a comparable error for all frames because the number of correspondences is not constant during the sequence. Until the registration error E is below an adaptive threshold, the algorithm proceeds with the alignment of the next frame starting from the camera parameters of the frame just aligned. This adaptive threshold is proportional to the objects surface area sampled by a single pixel of the camera. To be more precise, it is equal to the ratio between the width of the camera frustum at the distance of the object from the camera center and the width in pixels of the image. The distance of the object from the camera center is computed as the average between the maximum near and the minimum far plane of the camera to display only the portion of the object in the frustum. When the alignment error E is above the adaptive threshold, the algorithm applies the registration by Mutual Information.

After the registration by MI, the algorithm executes some additional steps to adjust the previous frames, in order to propagate backward the corrections of the camera parameters computed by the MI for the current frame. The algorithm re-

computes the correct 2D-3D correspondences of the current camera by backward projection of the 2D features. These correspondences are needed for the registration of the following frames. Subsequently, it updates all cameras between the current frame F_i and the last one F_{i-k} aligned by MI. For each camera in this interval it extracts the correspondences with the frames F_i and, for those cameras which have a minimum number of correspondences (at least 6), it recomputes new extrinsic parameters with the Levenberg-Marquardt algorithm using the 2D features shared with the frame F_i. Finally, for each of these frames it interpolates linearly the new extrinsic camera parameters with those computed before with the forward tracking, in order to obtain a continuous and smooth camera path without gaps.

The final task is the updating of the set of 2D-3D correspondences by backward projection of the new 2D features of the current frame that were not detected in the previous frame.

3.2.3 Registration by Mutual Information

Mutual Information measures the information shared by two random variables A and B. Mathematically, this can be expressed using entropy or joint probability. Following this interpretation, the Mutual Information $\mathcal{MI}$ between two images I_A and I_B can be defined as:

$$\mathcal{MI}(I_A, I_B) = \sum_{(a,b)} p(a,b) \log \left(\frac{p(a,b)}{p(a)p(b)} \right) \tag{3.2}$$

where $p(a,b)$ is the joint probability of the event that the same pixel in the images I_A and I_B gets the values a and b respectively, $p(a)$ is the probability that the pixel of I_A gets the value a and $p(b)$ is the probability that the pixel of I_B gets the value b (a comprehensive explanation of the Formula 3.2 can be found in [15]). The joint probability distribution can be estimated easily by evaluating the joint histogram ($\mathcal{H}$) of the two images and then dividing the number of occurrences of each entry by the total number of pixels. A joint histogram is a bi-dimensional histogram made up of $n \times n$ bins; the occurrence (a,b) is associated with the bin (i,j) where $i = \lfloor a/m \rfloor$ and $j = \lfloor b/m \rfloor$ and m is the width of the bin (Figure 3.3). The algorithm uses a joint histogram of 256×256 bins.

In the context of the registration, the extrinsic camera parameters are computed by maximizing the MI between the gradient of the frame and the gradient of a rendering of the 3D model with two illumination related properties, the surface normal and the ambient occlusion (Figure 3.4). This problem can be formalized as an optimization problem in a $6D$ space:

$$\begin{aligned} \mathcal{C}^* &= \arg\max_{\mathcal{C} \in \mathbb{R}^6} \mathcal{MI}(I_A, I_B(\mathcal{C})) \\ \mathcal{C} &= (t_x, t_y, t_z, \theta_x, \theta_y, \theta_z) \end{aligned} \tag{3.3}$$

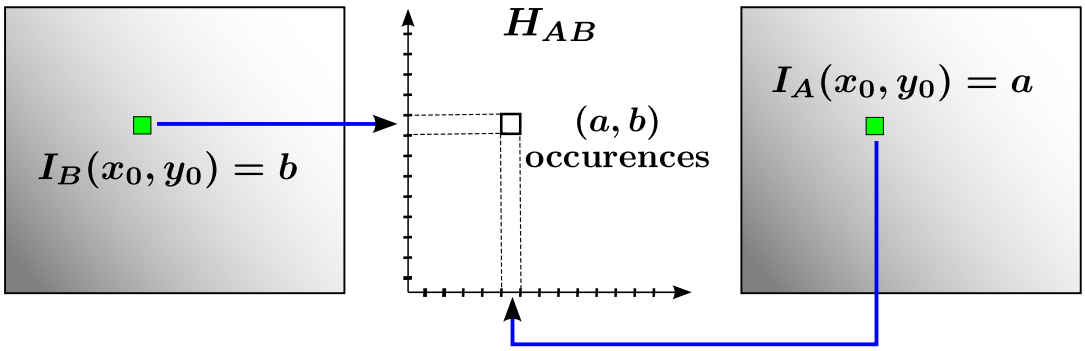


Figure 3.3: *Joint Histogram construction*

where (t_x, t_y, t_z) and $(\theta_x, \theta_y, \theta_z)$ define the position and orientation of the camera, I_A is the pre-processed image to align and I_B is the rendering of the 3D model. Hence, I_B depends on the camera parameters ($\mathcal{C}$).

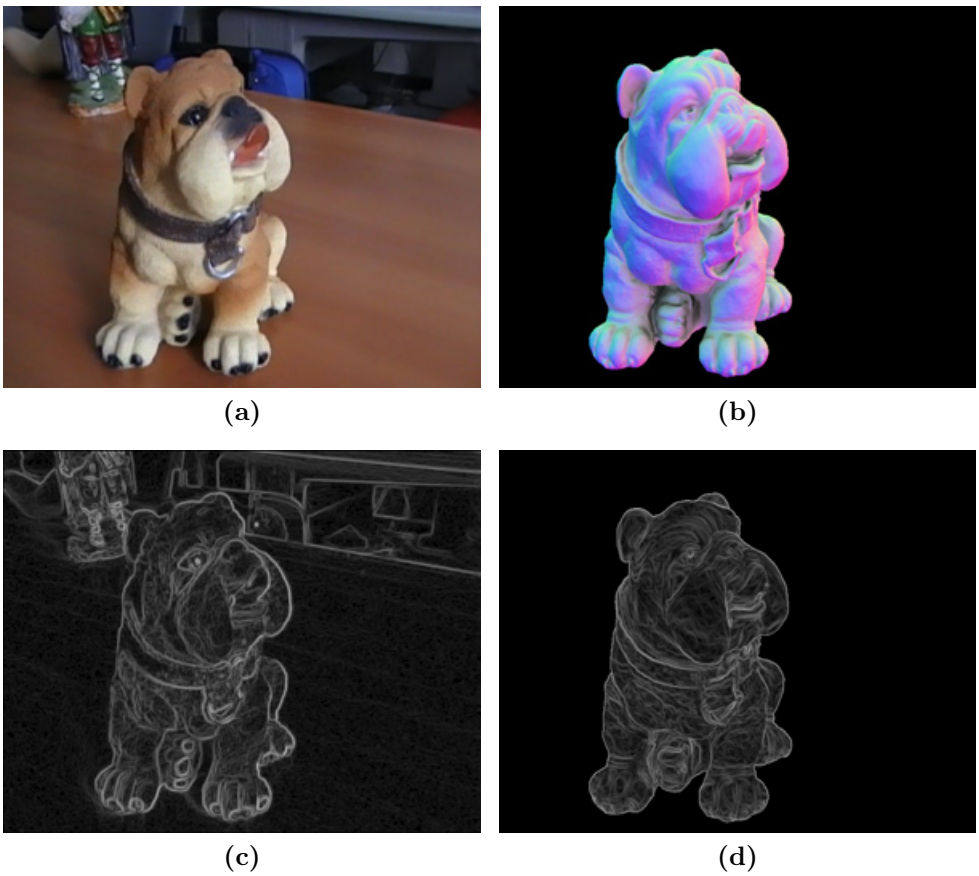


Figure 3.4: *Maps used to compute the registration by MI: (a) video frame; (b) rendering of the 3D model with a combination of normal map and ambient occlusion; (c) gradient map of the frame; (d) gradient map of the rendering.*

The algorithm extends the approach proposed in [34]. It generates a rendering of the 3D model with a combination of surface normal and ambient occlusion given the current camera parameters, it computes the gradient map of the rendering and the gradient map of the image and then it evaluates the mutual information between these gradient maps. An iterative optimization algorithm updates the camera parameters and recalculates MI until the registration is achieved. This means that the optimization algorithm reaches or a maximum number of iteration or a tolerance threshold. The image gradient is computed by applying the Sobel operator to the CIE luminance Y, computed using by the following formula:

$$Y = 0.2126R + 0.7152G + 0.0722B \tag{3.4}$$

In the computation of the joint histogram, the algorithm uses all the pixels in the rendering viewport, but it assigns a lower weight to the pixels on the background according the 3D rendering, in order to decrease their influence in the final registration.

The lack of a-priori knowledge about lighting, color and material reflectance information of the model prevents from generating realistic renderings. However, the goal of the rendering cycle is not to generate a photorealistic rendering but to synthesize an image that has a high correlation with the input picture under a wide range of lighting conditions and material appearances. On the other hand, the goal of the gradient is to maximize the shared data between the images decreasing the influence on the convergence toward the best camera parameters of two factors: the reflectance effects, like specular reflection and subsurface scattering, and the color features that the rendering cannot reproduce faithfully; the image background, especially when it is not uniform.

In general, the computation of the joint histogram on the gradient maps gives more importance to the geometric features with respect to the color features that the rendering cannot reproduce.

For the rendering of the 3D model the algorithm combines the information provided by the ambient occlusion and the normal map, as suggested in [34]. The ambient occlusion is precalculated and stored in the 3D model as per-vertex color using the algorithm proposed in [152]. During the rendering, the value of ambient occlusion is interpolated by Gouraud shading among the triangle vertices. The final color C is obtained by weighting the normal color map C_N with the value C_A of the ambient occlusion map (that is normalized between 0.0 and 1.0):

$$\begin{aligned} C_x &= (1 - C_A)C_A + C_A C_{Nx} \\ C_y &= (1 - C_A)C_A + C_A C_{Ny} \\ C_z &= (1 - C_A)C_A + C_A C_{Nz} \end{aligned} \tag{3.5}$$

where

$$C_N = \frac{\vec{N}}{2.0} + 0.5 \tag{3.6}$$

For the iterative optimization, the algorithm NEWUOA [160] is used. This algorithm iteratively minimizes a function $F(x)$, $x \in R^n$, by approximating it with a quadric Q in a region around the initial solution. A trust region procedure adjusts the variables looking for the minimum of Q, while new values of the function improve the approximation. The initial solution is the camera parameters computed using the 2D-3D correspondences in Levenberg-Marquardt optimization procedures.

A comparative study between the original registration approach [34] and the proposed extension with the gradient maps is presented in the Section 3.4.

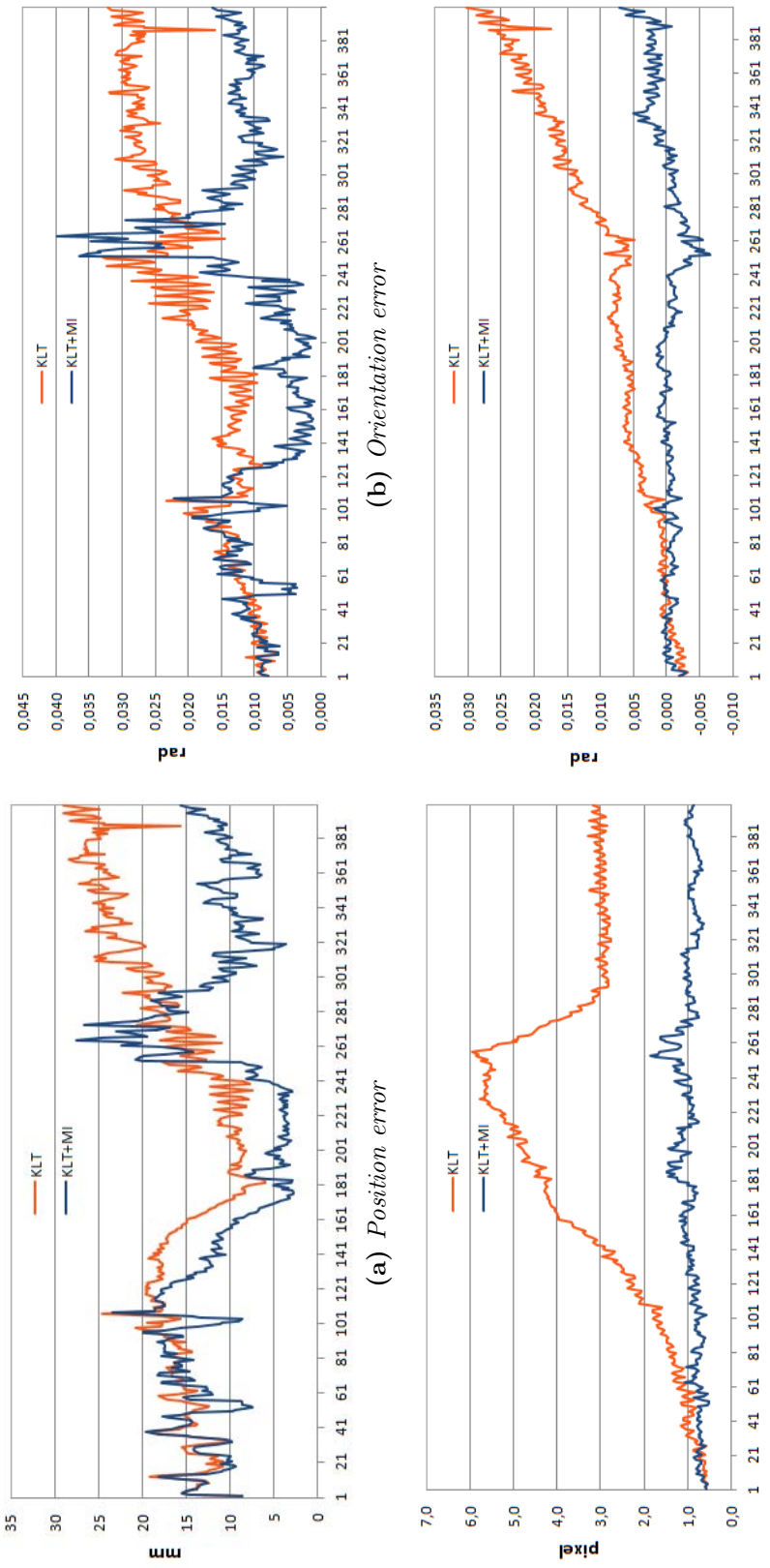
3.3 Video-to-Geometry Registration Results

This section presents and analyzes the results of the video-to-geometry registration for two different kinds of input sequences: a synthetic video, to evaluate the registration error and the effectiveness of the method quantitatively on a ground truth dataset, and a set of real video sequences of objects of different sizes. In order to evaluate the alignment improvements introduced by the registration with the Mutual Information, for each video sequence the camera parameters estimated with the proposed algorithm are compared with the parameters estimated using only the KLT tracking without the correction introduced with the Mutual Information. In order to visually evaluate the quality of the alignment, Figures 3.6, 3.7 and 3.8 show the original frame with an alpha blending of the rendering of the 3D model from the estimated camera.

3.3.1 Synthetic sequences

The synthetic video sequence is composed by 400 frames with known camera parameters and it is used to evaluate the quality and the precision of the registration achieved by the proposed algorithm. The evaluation compares the camera estimated by our method and the camera estimated only with KLT tracking data with respect to the ground truth. The sequence renders a colored 3D model (200k faces) of a medium height (50 cm) statue of a shepherd in a complex lighting environment, composed by an area light and an environment map. In the sequence, a set of possible effects, which characterize a real video sequence, are simulated, like motion blur, jittering, noise and unstable lighting conditions. The main causes of these effects are the type of camera motion and the environment and the camera characteristics, which break the quality of the KLT tracking.

Four different types of misalignment errors are computed and plotted in the charts in Figure 3.5. Each chart plots the error between the real camera and the camera estimated with our method (blue line) and with only the KLT tracking data (orange line). The chart 3.5a shows the distance in millimeters of the position of the estimated camera from the real one. The chart 3.5b shows the angle of the quaternion that defines the rotation needed to align the orientation of the estimated



(a) *Position error* **(b)** *Orientation error*

(c) *Projection error* **(d)** *Roll angle error*

Figure 3.5: *Charts of the registration errors: KLT + MI registration (blu line); KLT registration (orange line).*

camera with the real camera. The chart 3.5d shows the error in radiant of the roll angle of the camera around the optical axis. The chart 3.5c shows the projection error, which is computed by projecting a set of points uniformly distributed over the surface of the object in image space and calculating the average distance between the image points obtained by the real camera and the image points obtained by the estimated camera. The graphs show that the estimation of the cameras with the proposed method is better and less sensitive to the drift problem with respect to the camera recovered only with the KLT tracking data. This is particularly evident in the chart 3.5d. Another advantage of the algorithm is the very low and stable projection error (chart 3.5c). The analysis of the charts 3.5a and 3.5b requires more attention, especially between the frames 250 and 280. In this interval, the proposed method recovers a camera position and orientation with a higher error with respect to the camera estimate with only the tracking data, but the projection error is lower. This behavior is due to the statistical nature of the registration by MI that in this case converges towards a camera which is quite far away in space from the real camera, but very similar from the point of view of the projection as the chart 3.5c and the Figure 3.6 show.

3.3.2 Real-world sequences

The algorithm was also tested with four real sequences of different objects of known geometry: a dog small statue (about 20 centimeters tall); a shepherd statue (about 50 centimeters); a marble reproduction of an Ara-Pacis' bas-relief (about 2 meters); the Nettuno statue (about 6 meters) in the fountain on Piazza della Signoria in Florence. The object geometry was acquired by 3D scanning. The sequences were acquired with a consumer video camera with standard PAL resolution of 720×576 pixels and using a constant zoom factor. Figure 3.7 shows a visual comparison, for a specific frame, of the results obtained by the proposed registration algorithm and the results obtained using only the tracking data. A detail of the frame is shown to better visualize the misalignment. These results show the significant improvement introduced by the use of the MI, which corrects the drift error introduces by the incremental KLT tracking.

The results obtained in the sequence of the Nettuno statue are very interesting (Figure 3.8). In this sequence, a major occlusion appears during the video. The algorithm does not apply any strategy to discard the features that appear on the occluder object during the occlusion. As showed in Figure 3.8, the only use of the tracking data does not allow the estimation of a correct camera due to the incremental working of the registration that creates wrong 2D-3D correspondences during the occlusion. Consequentially, using only the tracking data, the algorithm loses a good registration, estimating wrong cameras for the frames during and after the occlusion. Conversely, the proposed algorithm preserves a good alignment even if the final registration is not very precise. In the specific, the algorithm estimates an unstable camera during the occlusion, but, in the subsequent frames, it is able to

Figure 3.6: *Registration results obtained in the synthetic sequence with KLT (Left) and KLT+MI (Right): frame 80 (Top); frame 264 (Center); detail of the frame 264 (Bottom).*

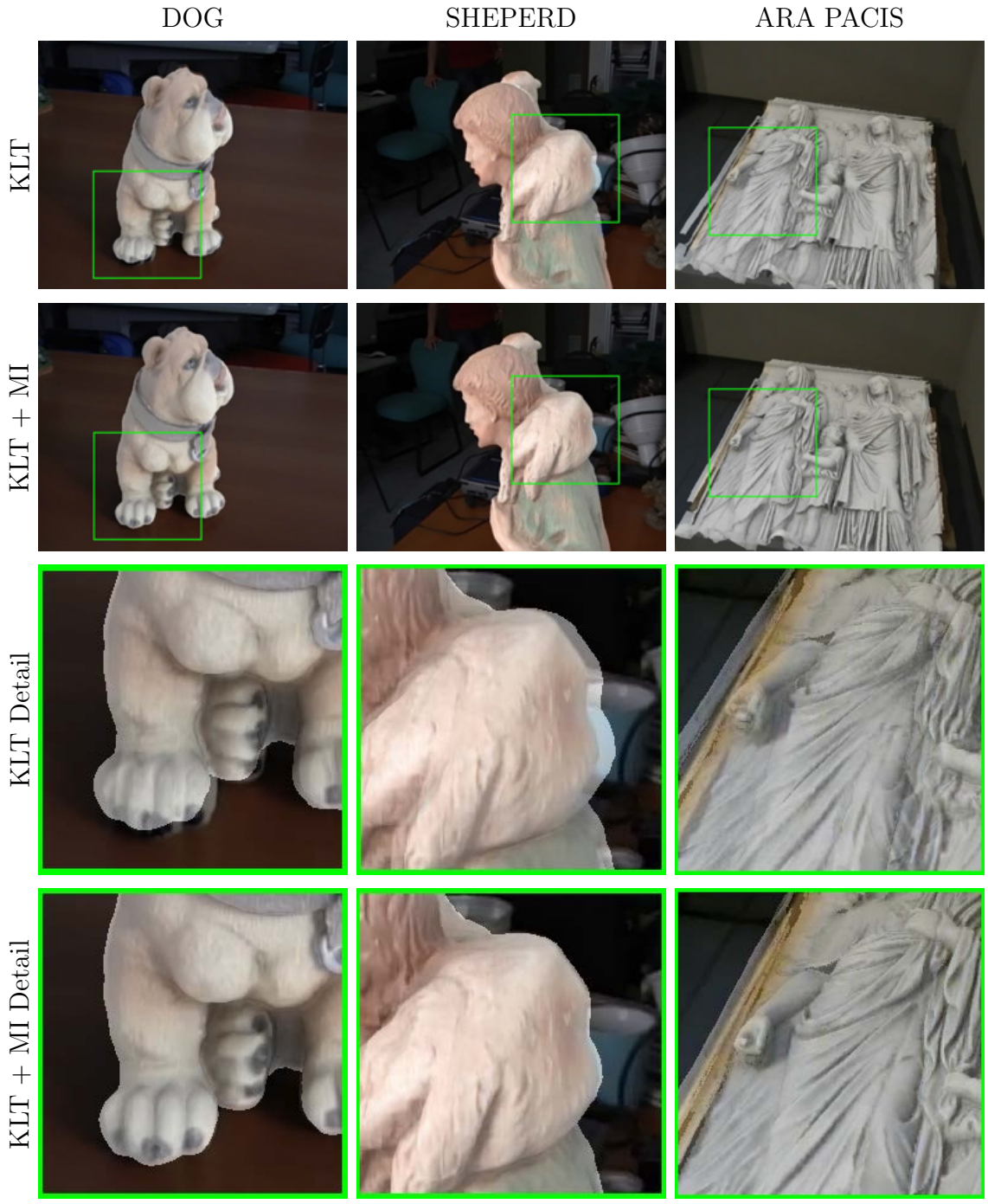


Figure 3.7: *Comparison of the registration obtained with real video sequences: Dog, frame 290 (Left); Shepherd, frame 400 (Center); Ara Pacis, frame 740 (Right).*

recover a good registration. In this challenging case, a further improvement in the precision of the registration during the occlusion can be obtained implementing a strategy to discard automatically the features on the occluders, for example taking into account the camera motion and the error information returned by the algorithm for each 2D-3D correspondence.

Figure 3.8: *Results on Nettuno sequence obtained by KLT registration (Left) and KLT+MI registration (Right): frame 20, before the occlusion (Top); frame 200, after the occlusion (Bottom).*

For each sequence, Table 3.1 shows the data to evaluate the performance of the algorithm: the length of the sequence; the number of triangles of 3D model used for the registration; the time required for the preprocessing of the video (deinterlace, denoise, removing of the lens distortion and tracking) and for the computation of the registration parameters; the number of frames on which the registration by MI is applied. The tests have been executed on a Intel Core2 Quad Q9400 with 4GB of RAM and a NVIDIA GTX260 896MB. The table highlights two aspects: the highest preprocessing time in the sequence of Shepherd's statue, due to the high number of features to track in the video, especially on the background; the highest registration time in the sequence of the Ara Pacis' bas-relief, due to the alignment

by MI that requires more iterations of the optimization algorithm NEWOUA to reach the convergence criteria.

	Frames	Geometry (triangles)	Preprocessing (mm:ss)	Registration (mm:ss)	MI (N. Frames)
Dog	347	195k	5:58	3:43	27
Shepherd	837	200k	16:18	11:43	73
Ara Pacis	749	350k	11:19	16:06	49
Nettuno	360	400k	7:16	5:56	81

Table 3.1: *Video-to-Geometry registration: performance data.*

3.4 Image-to-Geometry Registration using Gradient Maps

The main issue in the use of Mutual Information for 2D/3D registration is the choice of a rendering process that correlates the 3D model with the images to align. The main problem is that the input images contain texture and unknown lighting conditions: this could make their visual appearance very different from a rendering of the geometry. This section provides several experimental results in order to evaluate the improvements obtained by the use of the gradient map in the registration by Mutual Information, presented in Section 3.2.3, in a qualitative and quantitative way. The study compares the registration results obtained by the computation of the MI on the gradient maps with the results obtained by the framework proposed in [34], where the MI is computed directly on the rendering of the 3D model (normals + ambient occlusion) and on the image without computation of the gradient. In this comparison, the registration is formalized as an optimization problem in a 7D space, adding the possibility to estimate the focal length f (while the other intrinsic parameters are assumed as being pre-determined):

$$\begin{aligned} \mathcal{C}^* &= \arg\max_{\mathcal{C}\in\mathbb{R}^7} \mathcal{MI}(I_A, I_B(\mathcal{C})) \\ \mathcal{C} &= (t_x, t_y, t_z, \theta_x, \theta_y, \theta_z, f) \end{aligned} \tag{3.7}$$

The task is to evaluate the performance improvements obtained by the gradient maps analyzing for each test case the shape of the MI function around the optimal solution (Figures 3.9 3.10 3.11) and the convergence properties (Table 3.2). The test cases are five different objects, characterized by different reflection behaviors and geometric features. The photos were acquired with a digital camera with the exception of the DOG example, where a deinterlaced frame of a video acquired with a standard camcorder is used. All the photos were scaled to a width of 800 pixels to

have a comparable registration error among the test cases. The corresponding 3D models were generated by 3D scanning using a Konica Minolta VI910 laser scanner. For each test case the optimal registration was obtained using the semi-automatic tool TexAlign presented in [63], based on the selection of 2D-3D correspondences to use in the Tsai's calibration method [192]. The error in the optimal solution is estimated to be about one pixel.

		Convergence (% of success)				
Test	Map	Initial perturbation (pixels)				
		10	20	30	40	50
HORSE	Gradient	100	100	91	83	75
	Norm+Amb [34]	100	95	84	75	34
SHEPHERD	Gradient	100	95	88	70	55
	Norm+Amb [34]	57	70	70	51	46
DOG	Gradient	80	88	60	22	12
	Norm+Amb [34]	18	7	9	3	1
OMOTONDO	Gradient	100	49	35	12	4
	Norm+Amb [34]	50	22	16	8	5
GARGOYLE	Gradient	100	98	94	88	86
	Norm+Amb [34]	100	91	36	10	4

Table 3.2: *Registration by MI using gradient maps: convergence tests.*

The shape of the MI function is drawn by evaluating the function in the neighborhood of the optimal solution. Since the MI function around the aligned position is a function of seven camera parameters, the overall shape around the aligned position is explored with a number of 1D sections (30 sections in the plot in Figures 3.9, 3.10 and 3.11) calculated in random directions in the 7D space. In this way in correspondence to a local minimum, every 1D section should exhibit the same minimum. Figures 3.9, 3.10 and 3.11 show a comparison between the MI function graphs of the proposed algorithm (right column) and the graphs of the method by Corsini et al. [34] (central column) using normal map and ambient occlusion. In the specific, the quality of the MI function is defined by its shape: the most important factors are the existence of a well defined minimum and a smooth and monotonic shape around the minimum, which permits a wider range of convergence. Analyzing the graphs, it is possible to conclude that the use of the gradient allows the generation of smoother function with better convergence properties near the minimum due to a higher curvature. Especially for the three examples DOG, OMOTONDO and GARGOYLE the improvement is more evident.

The convergence properties of the algorithm are tested by applying 300 random

perturbations to the camera parameters of the optimal solution and then measuring the convergence rate of the algorithm. The parameters were perturbed simultaneously with several amount of perturbations that are measured in image space (10, 20, 30, 40, 50 pixels). Starting from a set of uniform distributed 3D points, the perturbation is measured as average of the Euclidean distance between the image point obtained by projection with the perturbed camera and the image point obtained with the reference camera. The maximum allowable perturbation with respect to the reference registration was 50 pixels. For each perturbation value the convergence rate is measured as the percentage of success in convergence of the algorithm, defined as the number of times that the final registration error is less than 2 pixels with respect to the ground truth obtained by the TexAlign tool. Table 3.2 shows that the convergence percentage obtained with the gradient maps is higher. Generally, for large perturbations, like 40 or 50 pixels, the difference between the convergence rates becomes more marked. Especially in the DOG example, where a camcorder is used, the big improvement introduced by the gradient maps decreases the influence of the background and of the degrading factors of the acquisition system, such as noise and lens distortions. A general improvement for all the perturbation values are obtained in the SHEPHERD example, where the image is acquired with a spotlight, and in the OMOTONDO example, while the HORSE and GARGOYLE examples show the most evident improvement in the convergence rate for large perturbations (40 and 50 pixels).

Figure 3.9: *MI function plots for the HORSE and SHEPHERD test cases: (Left Column) photo; (Central Column) MI function graphs for Normal+Ambient Occlusion rendering proposed in [34]; (Right Column) MI function graphs for Gradient Map.*

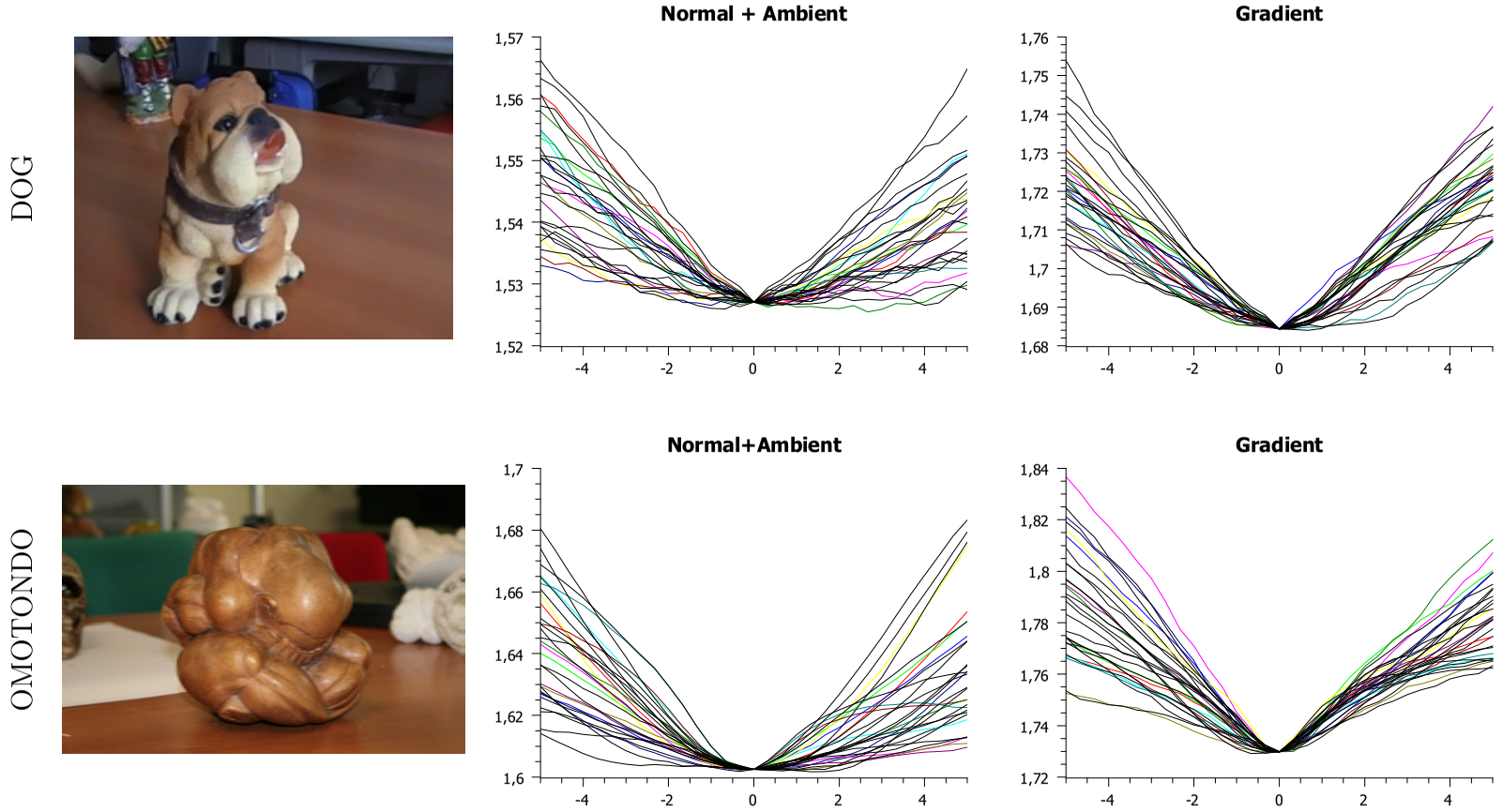


Figure 3.10: *MI function plots for the DOG and OMOTONDO test cases: (Left Column) photo; (Central Column) MI function graphs for Normal+Ambient Occlusion rendering proposed in [34]; (Right Column) MI function graphs for Gradient Map.*

Figure 3.11: *MI function plots for the GARGOYLE test case: (Left Column) photo; (Central Column) MI function graphs for Normal+Ambient Occlusion rendering proposed in [34]; (Right Column) MI function graphs for Gradient Map.*

Chapter 4

Surface Light Field from Video Made Easy

This chapter presents an algorithm for the estimation of the Surface Light Field using video sequences acquired moving the camera around the object. Unlike other state of the art methods, it does not require a uniform sampling density of the view direction allowing to reconstruct an estimation even from a biased video acquisition: very dense only along the camera path and completely missing in the other directions. The main idea is to separate the estimation of two components: the diffuse color and the other residual Surface Light Field effects. The diffuse component is computed using statistical operations that allow the estimation of the rough direction of the main light sources, which were in the acquisition environment. This approximation of the lighting environment is used to discard the color samples that have a higher probability to exhibit view-dependent reflectance effects. The residual component is modeled as linear combination of spherical functions (Spherical and Hemispherical Harmonics). From a qualitative and numerical evaluation, the final rendering results show a high fidelity and similarity with the input video frames, without artifacts due for example to the trivial fitting of the acquired samples on a set of spherical functions, which can produce ringing and banding effects.

4.1 Surface Light Field Estimation

The appearance acquisition and estimation is an open problem in the CH due to the reflectance characteristics of the object and the constrains imposed by the acquisition environment. A CH artwork can be composed by different materials, which could present several types of patinas and degradation processes that alter the appearance, making difficult its modeling. Usually artworks cannot be moved in an acquisition lab and the lighting environment cannot be substantially modified, imposing to perform on-the-field acquisitions without the possibility to used controlled lighting conditions, like a dark room. Even if solutions have been proposed to estimate

more complex appearance model in these challenging conditions (a representative is the method proposed in the Chapter 5), in some applications it is sufficient to estimate a further simplification of the appearance in an even more automatic way. The typical example is a virtual museum, where, in order to reproduce a real visit, the artwork can be viewed by different positions but with fixed lighting conditions. In this contest, an appropriate solution is the Light Field Rendering [112] and its following extensions, image-based rendering approach that uses sets of photos, taken from different points of view, to be resampled at rendering time to generate novel images. One of the major benefits of these techniques is the ability to create realistic renderings of a wide range of physical surfaces (including anisotropic ones) with complex reflectance behavior, without passing through the complex definition and choice of intricate reflection models. Nevertheless, image-based techniques have some drawbacks in term of data complexity if the accuracy and the quality of the final rendering must be increased, due to the huge amount of data to acquire. For this reason hybrid solutions have been proposed. The main idea is to combine acquired photos with some geometric information to synthesize new views of the scene. The amount of geometric information employed varies from depth maps to reasonably detailed polygonal meshes. Among the proposed solutions, there are the Lumigraph [78] and the Surface Light Field (SLF) [134].

An important issue in image-based and hybrid methods for the Light Field estimation is the sampling density of the view directions. Many state of the art algorithms require a dense and uniform photographic acquisition of the object, obtained for example with special devices like camera arrays or robotic arms. However, in order to make the method usable in the Cultural Heritage context, it is necessary an automatic system that is able to reconstruct the appearance from partial and irregular acquisitions, reducing the expertise of the operator needed to evaluate if the acquired data are enough for the reconstruction. Acquired data are often partial because of the need to perform a fast sampling (not supported by a careful planning), the physical limitations and constraints of the digitization environment (for example the presence of obstacles and occluders), or the challenging position and size of the artwork, which limit the free movement of the camera around the object and prevent to capture a complete viewpoint sampling. This chapter is focused on the challenging problem of the Light Field estimation from a not uniform and irregular spatially distributed acquisition, like a video sequence acquired by simply moving the camera around the object.

The chapter presents a method for the estimation of the Surface Light Field of a real-world object. It starts from a medium quality 3D model and some video sequences, which do not guarantee a uniform sampling of the view direction. The typical video acquisition is composed by very simple movements around the object, returning a view sampling biased by the camera movement: very dense only along the camera path and completely missing in the other directions. Figure 4.1 shows some typical acquisition path with this feature while Figure 4.2 shows the distribution of the acquired color samples in the visible hemisphere for a single surface point. The

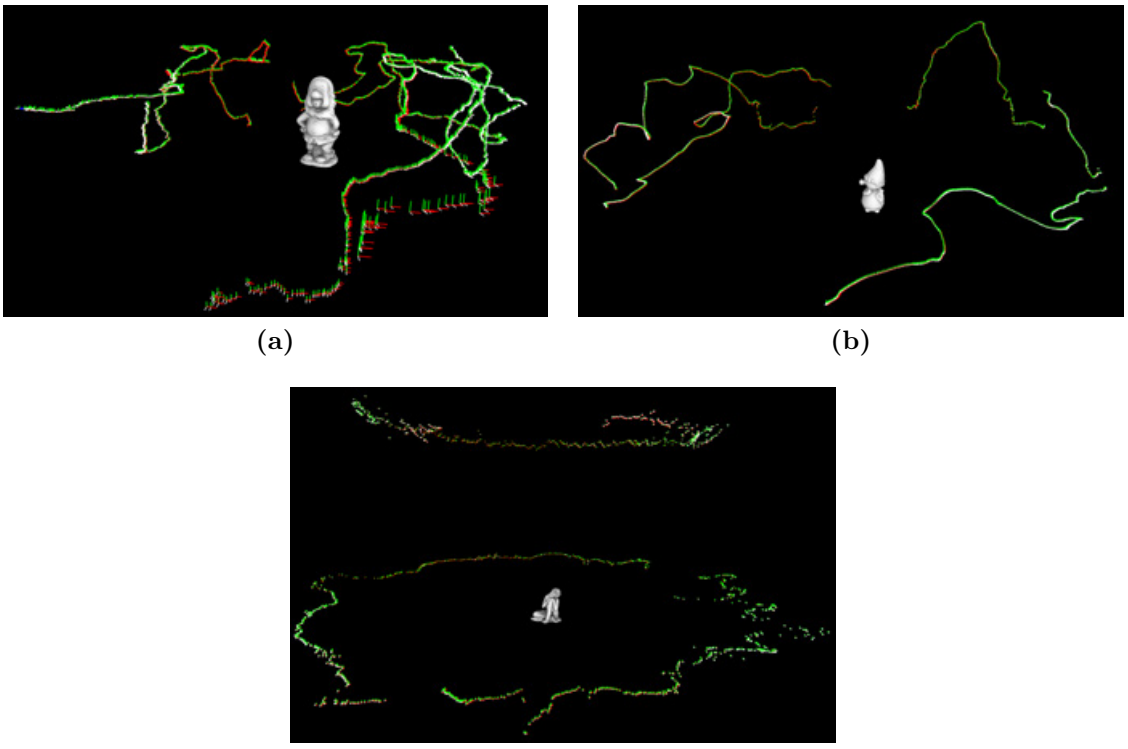


Figure 4.1: *Test cases camera paths: (a) DWARF; (b) GNOME; (c) SLEEPING BUDDHA*

algorithm uses the idea of the Dichromatic Reflection Model [175] to separate the estimation of the diffuse color from the estimation of the other surface appearance effects. For the diffuse color, it uses a statistical method, based on the selection of the samples with a higher probability to have a Lambertian behavior, while the residual color, which is the difference from the diffuse color, is used to fit a linear interpolation of spherical functions (Spherical and Hemispherical Harmonics).

Main contributions of the work are:

- a method that, starting from a video acquisition of the object with a not uniform viewpoint sampling, is able to estimate the Surface Light Field without banding and ringing artifacts due to the spherical function fitting and the irregular distribution of the input color samples in the visible hemisphere (examples of these artifacts are shown in Figure 4.3 with a comparison of the results obtained by the method presented in this chapter);

- a method able to extract the Lambertian shading of the object using statistical operations that simplifies the modeling of the other Surface Light Field effects.

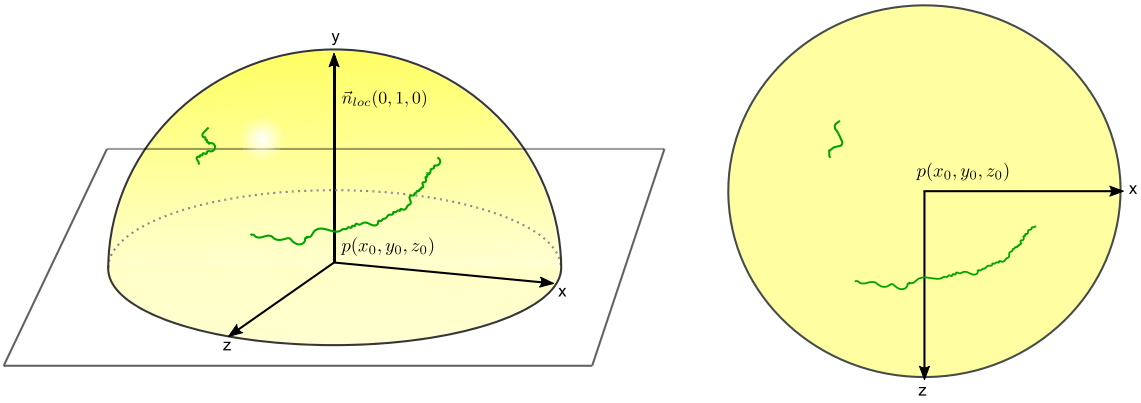


Figure 4.2: *Typical distribution of our input data in the visible hemisphere for a single surface point: (Left) Visible hemisphere; (Right) Top orthogonal projection. The green points are the view directions acquired by the video sequences.*

Even if the proposed method is not able to accurately predict the view dependent effects from not acquired viewpoint, due to the irregular acquisition, the main goal is to obtain plausible renderings without artifacts (Figure 4.3).

4.1.1 Background

The Surface Light Field was introduced in [134] to sample Light Field on a parametric surface directly. It is modeled as a four-dimensional function $L(u, v, s, t)$, where (u, v) defines a point on the surface and (s, t) represents the orientation angles identifying the view direction (the spherical coordinates of the view vector with respect to the surface normal). Given a novel camera position, the rendering is done by computing the view direction values (s, t) for each surface point (u, v) and then indexing the Surface Light Field using (u, v, s, t) to extract the color value. The use of the geometry of the object allows overcoming some of the drawbacks of a generic Light Field representation: the camera is restricted to certain regions of space; the finite angular resolution leads to depth of field effects; objects become blurry in proportion to their distance from the image plane. The Surface Light Field representation allows the rendering of images with view-dependent effects, for example highly specular objects in arbitrary complex lighting environments, with a good and faithful reproduction of interreflections and shadows.

The proposed method is based on the separation of the diffuse component of the surface appearance of the object from the other view dependent lighting effects. This separation process avoids rendering artifact due to the fitting and interpolation process of the spherical functions. This idea is mathematically supported by the Dichromatic Reflection Model [175]. Intuitively, this model states that there are two independent reflection processes, the specular and the diffuse reflectance, each one characterized by a color whose magnitude varies with the light $\vec{l}$ and view $\vec{v}$

Figure 4.3: *Surface Light Field rendering from viewpoints not acquired by the video sequences: (Left) results from the trivial fitting of the input samples with the first 25 Spherical Harmonics functions; (Right) results of the proposed algorithm. The trivial fitting produces artifacts, like ringing and banding effects, for two reasons: the irregular viewpoint sampling and the missing color samples in wide areas of the visible hemisphere.*

directions:

$$C(\vec{l},\vec{v}) = m_d(\vec{l})C_d + m_s(\vec{l},\vec{v})C_s \tag{4.1}$$

where C is the final RGB color, C_d and C_s the color of the two components, and m_d and m_s are the magnitude of the two components. However, in the context of Surface Light Fields, since there are fixed lighting conditions and assuming to have a Lambertian diffuse component, the Equation 4.1 can be simplified:

$$C(\vec{v}) = C_d + m_s(\vec{v})C_s \tag{4.2}$$

4.2 The Algorithm

The algorithm takes in input a medium-resolution triangular mesh of the object, with an associated texture parameterization, and some video sequences, acquired moving the camera around the object. The aim is to estimate the diffuse color and the other view dependent Surface Light Field effects in two different steps. In the first one, the algorithm estimates the diffuse component using statistical operations. In the second step, it reconstructs the other Surface Light Field effects (mainly specular effects) as linear combination of spherical functions. The final color of a point p with texture coordinates u, v is given by the following formula:

$$C(u,v,s,t) = D^{(u,v)} + \sum_{i=0}^{n} x_i^{(u,v)} h_i(s,t) \tag{4.3}$$

where (s,t) are the spherical coordinate (θ,ϕ) of the view vector with respect to the surface normal in p, $D^{(u,v)}$ is the diffuse texture color of the point, $x_i^{(u,v)}$ are the coefficients that are associated to the selected basis of spherical functions $h_i(s,t)$. The outputs of the algorithm are a texture map with the diffuse color $D^{(u,v)}$ and a binary file with the coefficients $x_i^{(u,v)}$, organized as an OpenGL texture array, where the i-th layers contains the coefficient $x_i^{(u,v)}$.

The estimation process is organized in four steps:

1. the video-to-geometry registration (Section 4.2.1);

2. the estimation of the direction of the main light sources in the acquisition environment (Section 4.2.2);

3. the approximation of the surface diffuse color (Section 4.2.3);

4. the fitting of the view dependent Surface Light Field effects (as residual from the diffuse color) in a set of basis spherical functions (Section 4.2.4).

4.2.1 Video-to-Geometry Registration

Using the method proposed in the Chapter 3, the intrinsic and extrinsic camera parameters are recovered for each frame in order to align the video on the 3D model. The result of the registration process allows the computation of the color samples $C_{u,v} = \{I^{(j)}_{u,v} \in RGB\}$ and the relative quality value $q^{(j)}_{u,v}$ projected by each frame j for each texel (u, v). The quality value is computed using the framework proposed in [23]. It is equal to the product of three measures: the distance in image space from the nearest depth discontinuity (Figure 4.4a), to penalize possible wrong color samples due to small misalignment; the depth of the texel in camera space (Figure 4.4b), to give a higher quality at the samples acquired by closer views; the view angle (Figure 4.4c), to give a higher quality at the samples that have a small angle from the view direction.

(a) **(b)** **(c)** **(d)**

Figure 4.4: *Per pixel quality function: (a) map of the border distance from the depth discontinuities; (b) depth map; (c) dot product between the normal and the view direction; (d) final quality.*

4.2.2 Light Direction Estimation

In this step, the main goal is the approximation of the direction of the main light sources that were in the environment during the acquisition, assuming distance light sources. It starts with the construction of a rough environment map using a simple accumulation approach along the specular mirror direction, followed by a clustering method to isolate the main light sources and compute their main direction.

In the specific, for each texel (u, v) independently, it selects all the projected color samples that have a higher probability to show a not diffuse behavior. It uses a trivial approach based on luminance thresholding, which selects all the color samples near or in the saturation area of the camera sensor. In the specific it selects all the samples with a luminance greater than a fixed threshold $t = 0.98$ (with a luminance range $[0.0, 1.0]$). For each selected sample, it computes the specular mirror direction

$\vec{r}$ of the view vector $\vec{v}$, and it increments the value of the pixel in the environment map where the spherical coordinates of the vector $\vec{r}$ are mapped. Then, the map is normalized with the distribution of all color samples in the environment along the specular mirror direction. This distribution is computed as the total number of samples that project in each specific pixel of the environment map along the direction $\vec{r}$. This normalization gives more robustness with respect to the camera movement, especially when the temporal density of the acquisition is not uniform on the whole object (for example very fast in front of the object and slower in the back). Finally, the environment map is normalized in the range $[0, 1]$ with respect to the maximum value.

Even if the trivial luminance thresholding with a fixed threshold is prone to introduce noise in the estimation of the environment map, especially for white object, the high redundancy of the video data makes the estimation process more robust and stable. The aim of this threshold is to increase the statical influence of the samples with a higher probability to show a specular behavior, which is the behavior required by an optimal light probe.

(a) **(b)**

Figure 4.5: *Comparison of the estimated acquisition environment map (b) with the real one obtained from a photo of a reflective metal sphere (a).*

An example of the estimated environment map, with a comparison with the real one acquired with a reflective metal sphere, is shown in Figure 4.5. The reconstructed map (Figure 4.5b) gives some clues about the position of the main light sources in the scene, creating several clusters that correspond to the main light sources of the real environment map (Figure 4.5a). To detect these clusters the algorithm uses a K-Mean clustering, where the number of clusters can be selected

by the user or automatically by finding the local maximums in the environment. Finally it computes the centroid of each cluster to obtain an estimation of the light direction $\vec{l}_k$.

4.2.3 Diffuse Color Approximation

The light directions computed in the Section 4.2.2 are used to make more robust the estimation of the diffuse color. The computation is performed in three steps, for each texel independently. The main idea is to take advantage of the high amount of color samples projected on each texel to discard all the samples that have a higher probability to introduce inaccuracy in the computation of the diffuse color, such as specular samples or wrong projected samples near color discontinuities due to small registration errors.

In the first step, color samples are sorted by their luminance value to discard a percentage $p_t = 15\%$ of them having the lowest luminance. This step increases the local coherence near the color boundaries, avoiding abrupt changes in luminance due to wrong projected color.

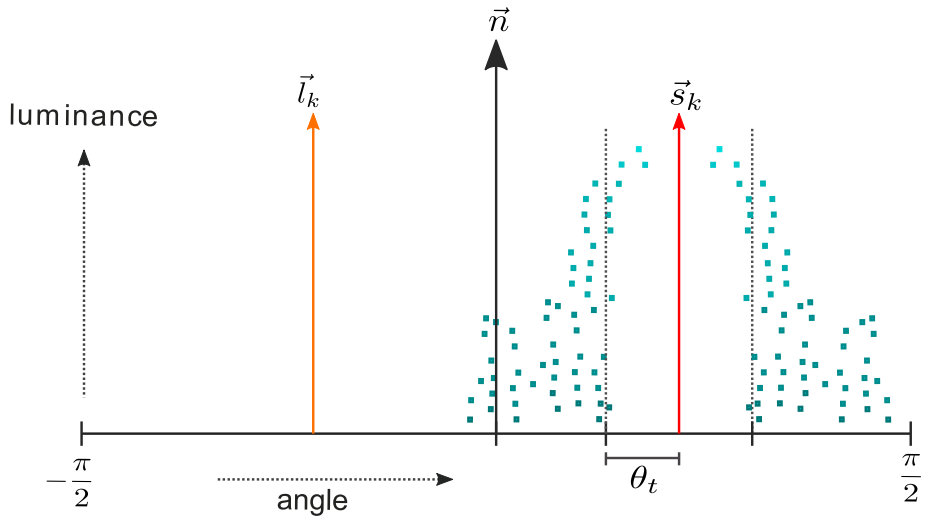


Figure 4.6: *Graph of the distribution of the samples with respect to the angle between the view direction and the surface normal. All the samples with an angle from the specular mirror direction $\vec{s}_k$ of the light vector $\vec{l}_k$ less than the threshold θ_t are discarded in the estimation of the diffuse color*

.

In the second step, it discards all the color samples with a higher probability to have a view-dependent reflectance behavior, especially specular reflection. Knowing that specularity occurs when the view vector is aligned with the specular mirror direction of the light vector, the idea is to determine which samples appear close

to this direction and discard them. For this purpose, each light $\vec{l}_k$ is reflected with respect to the surface normal $\vec{n}$ to compute the specular mirror direction $\vec{s}_k$. Then the algorithm discards the color samples with an angle between its view direction $\vec{v}$ and the direction $\vec{s}_k$ less than a given threshold $\theta_t = \pi/12$, as depicted in Figure 4.6. The purpose of the thresholds p_t and θ_t is not to discard all the samples with bad properties for the estimation of the diffuse color, but to decrease the number of them in order to reduce their influence in the following statistical computation of the diffuse color. They are chosen to work in the most general cases. This means that the two thresholds do not guarantee to discard all and only the specular samples, but the redundancy and the temporal coherence of the videos allow the achievement of good results.

At this point, the set $\widetilde{S}_{u,v}$ of the remaining samples is used to compute a threshold $t_{u,v}$ that is an upper bound of the median:

$$t_{u,v} = \mu_{u,v} + \bar{\sigma}_{u,v} \tag{4.4}$$

where $\mu_{u,v}$ and $\bar{\sigma}_{u,v}$ are the mean and the average absolute deviation of the luminance of the samples in $\widetilde{S}_{u,v}$, weighted with the quality $q^{(j)}_{u,v}$. Finally, the diffuse color $D^{u,v}$ is computed as weighted average with the quality $q^{(j)}_{u,v}$ of the color samples that have a luminance value lower than the threshold $t_{u,v}$.

4.2.4 Color Residual Fitting

The idea is to use the residual images, obtained as positive difference of the color samples from the diffuse color, to compute the coefficients x_i of a linear combination of a basis of spherical functions h_i, which just depends on the local spherical coordinates (s, t) of the the view direction:

$$\sum_{i=0}^{n} x_i^{(u,v)} h_i(s,t) \tag{4.5}$$

The algorithm was tested with two different basis of spherical functions: Spherical Harmonics and Hemispherical Harmonics [67]. Using the Dichromatic Reflection Model and assuming a white light, the color residual is modeled using only the luminance difference from the diffuse color.

For each texel (u, v) it retrieves the set of color samples $S_{u,v}$ that have a positive residual from the diffuse color and it sets a system of linear equations:

$$A\mathbf{x} = \mathbf{b} \tag{4.6}$$

where A is an $m \times n$ matrix that for each row, one for each sample in $S_{u,v}$, contains the values of the spherical functions computed for the view direction of the sample, $\mathbf{x}$ is the vector of the n coefficients x_i to estimate, and $\mathbf{b}$ is the vector with the luminance difference from the diffuse color.

To solve the overdetermined system in Equation 4.6 the algorithm uses a Weighted Singular Value Decomposition (SVD) in order to take advantage of the quality information $q_{u,v}^{(j)}$ related to each sample. In this way, it computes a weighted least square solution of the system:

$$\underset{\{x_1,\ldots,x_n\}}{\arg\min} \left(\sum_{j=1}^{m} q^{(j)} r_j^2 \right) \tag{4.7}$$

where

$$r_j = \left(\sum_{i=0}^{n} x_i \; h_i(s_j, t_j) \right) - b_j \tag{4.8}$$

The weighted least square solution reduces the influence of a combination of misalignment of the videos on the geometry and artifacts in the video frames, which can alter the realism of the result. In general the samples in $S_{u,v}$, used to solve the equation 4.7, cover only a small part of the visible hemisphere (see Figure 4.2). To avoid that the fitting procedure creates artifacts in the not sampled areas, the algorithm adds some virtual samples, uniformly distributed in the uncovered regions, with a residual color equal to zero (at most 400 samples distributed with a Poisson-Disk strategies with respect to the existing samples). In this way, it avoids artifacts due to banding and ringing effects.

Finally to reduce the GPU memory footprint need for the rendering of the Surface Light Field, the floating-point coefficients x_i are compressed with a simple 8-bit quantization. For each coefficient it finds the maximum and minimum values $(\text{MAX}_i, \text{MIN}_i)$ and it computes a scale λ_i and bias Ω_i factors:

$$\begin{aligned} \lambda_i &= \frac{\text{MAX}_i - \text{MIN}_i}{255} \\ \Omega_i &= \text{MIN}_i \end{aligned} \tag{4.9}$$

that are used during the rendering with the compressed coefficients $\widehat{x_i}$ to recover the original coefficients x_i:

$$x_i = \widehat{x}_i \lambda_i + \Omega_i \tag{4.10}$$

4.3 Results

Three different objects of different materials are used to test the algorithm:

- the DWARF, a terracotta statue (30cm tall) that presents different types of specularity, in size and intensity: sharper and with a high-medium intensity on the dress; wider on the face; almost completely absent on the beard;
- the GNOME, a ceramic statue (15cm tall) that has very sharp and high specularity on the hat and a near diffuse behavior on the body;

	Geometry (triangles)	Mesh Processing (minutes)	Frames	Registration time (minutes)	SLF estimation (minutes)
DWARF	200k	55	3382	113	72
GNOME	135k	40	2092	73	61
SLEEPING BUDDHA	205k	70	2414	83	65

Table 4.1: *Surface Light Field Estimation: Models and input datasets characterization*

- the SLEEPING BUDDHA, an acrylic resin Buddha (10cm tall) with different types of coatings (a gold paint on the body, a reddish specular paint on the dress and a diffuse black paint on the hair).

The videos were acquired with a full HD video camera that are set at the highest acquisition quality to reduce the compression artifacts, while the 3D models were generated by 3D laser scanning and then simplified to obtain a medium resolution model for the computation of the texture parameterization. For each object the algorithm uses a 2048×2048 texture. All the data about the datasets are shown in Table 4.1 (the size in triangles of the 3D model, the time for the generation of the medium resolution 3D model, the length in frames of the videos used for the estimation of the SLF, the time required for the alignment of the video on the mesh using the method presented in the Chapter 3, the time required for the computation of our approximation of the SLF). The tests have been executed on a PC with an Intel Core i7 950 with 12GB of RAM and a NVIDIA GTX580 1536MB

The figures 4.7, 4.8 and 4.9 show a comparison of the results of the proposed method with an original frame of the video used by the algorithm. In the specific for each figure, they show the original frame (a), the estimation results of the diffuse color (c), and the estimation results of the Surface Light Field using two different basis of spherical functions: Hemispherical Harmonics (HSH) in the sub-Figures (d), (e) and (f); Spherical Harmonics (SH) in the sub-figures (g), (h) and (i). In the Surface Light Field estimation an increasing number of coefficients are tested (4, 9 and 16 coefficients for HSH and 9, 16 and 25 coefficients for SH), up to a number that allows the real-time rendering (above 25 fps) of the obtained Surface Light Field on the last common GPUs. From a qualitative and visual evaluation of the results, its possible to conclude that the method proposed in the Section 4.2.3 is able to estimate a good approximation of the Lambertian shading of the object without artifacts and discontinuities, removing the majority of the other reflection effects, like the specularities (compare the Figures 4.7a, 4.8a and 4.9a with the Figures 4.7c, 4.8c and 4.9c). Furthermore, the final SLF has a good similarity with the original frame, a similarity that increases with the number of used spherical functions.

To have a more objective evaluation of the obtained results, two different metrics to measure the image fidelity between the original frames and the obtained renderings are used: the Mean Squared Error (MSE); the Structural SIMilarity (SSIM)

	DWARF		GNOME		SLEPPING BUDDHA	
	MSE	SSIM	MSE	SSIM	MSE	SSIM
Diffuse	0.003075	85.187%	0.003089	76.533%	0.005666	79.546%
HSH - 2 band	0.002027	87.169%	0.003059	76.879%	0.002728	83.788%
HSH - 3 band	0.001259	88.365%	0.002661	78.211%	0.001972	85.948%
HSH - 4 band	0.001020	89.040%	0.002543	79.024%	0.001538	87.322%
SH - 3 band	0.001317	87.905%	0.002695	78.099%	0.002237	84.402%
SH - 4 band	0.001066	88.478%	0.002581	78.673%	0.001788	85.650%
SH - 5 band	0.000932	88.818%	0.002515	79.232%	0.001635	86.178%
SH Enhanced	0.000869	91.414%	0.002473	82.998%	0.001081	89.481%

Table 4.2: *Error measures*

index [201], a perceived fidelity measure (the value is a percentage and the value 100% is returned when an image is compared with itself). Table 4.2 contains the value of these metrics for the three test cases. The table includes as well the error between the original images and the renderings with only the diffuse component (Figures 4.7c, 4.8c and 4.9c), in order to highlight the improvement introduced by the term that models the residual color. The data in the table confirms how the accuracy and fidelity of the SLF increase with the number of used spherical functions and, given a number of coefficients, the HSH representation outperforms the SH representation.

The renderings show some small differences from the reference frames: in the diffuse color the algorithm loses some very small details due to small misalignment in the video-to-geometry registration; some highlights appear different from the reference frame due mainly to imprecise normals (for example on the hat of the GNOME); the proposed method does not reproduce some specular highlights due to Fresnel effects (for example on the nose and on the top-right of the face of the DWARF) because this type of effect appears near to the silhouette of the object, where the algorithm gives a lower quality at the samples; finally the highlights appear less bright than the original ones in the original frames due to the limited number of functions (at most 25 functions for SH and 16 for HSH) used to model the residual color. To be more specific, the limited number of spherical functions used does not allow the reproduction of the narrower specularity peaks, obtaining a band-limited reconstruction. This aspect is confirmed by the higher rendering quality obtained by increasing the number of functions. The use of a higher band of spherical approximation can surely remove this type of imperfections but at the cost of lower rendering performance. In order to guarantee the real-time rendering of the proposed SLF representation, a further improvement of the visual results can be obtained by introducing a small change in the rendering Equation 4.3, based on the enhancement of the residual component. The equation is changed by adding a new term I_s that is used in the following manner to vary the intensity of the residual

(a) *Original frame* **(b)** *Enhanced SLF* **(c)** *Diffuse Color*

(d) *HSH - 2 bands* **(e)** *HSH - 3 bands* **(f)** *HSH - 4 bands*

(g) *SH - 3 bands* **(h)** *SH - 4 bands* **(i)** *SH - 5 bands*

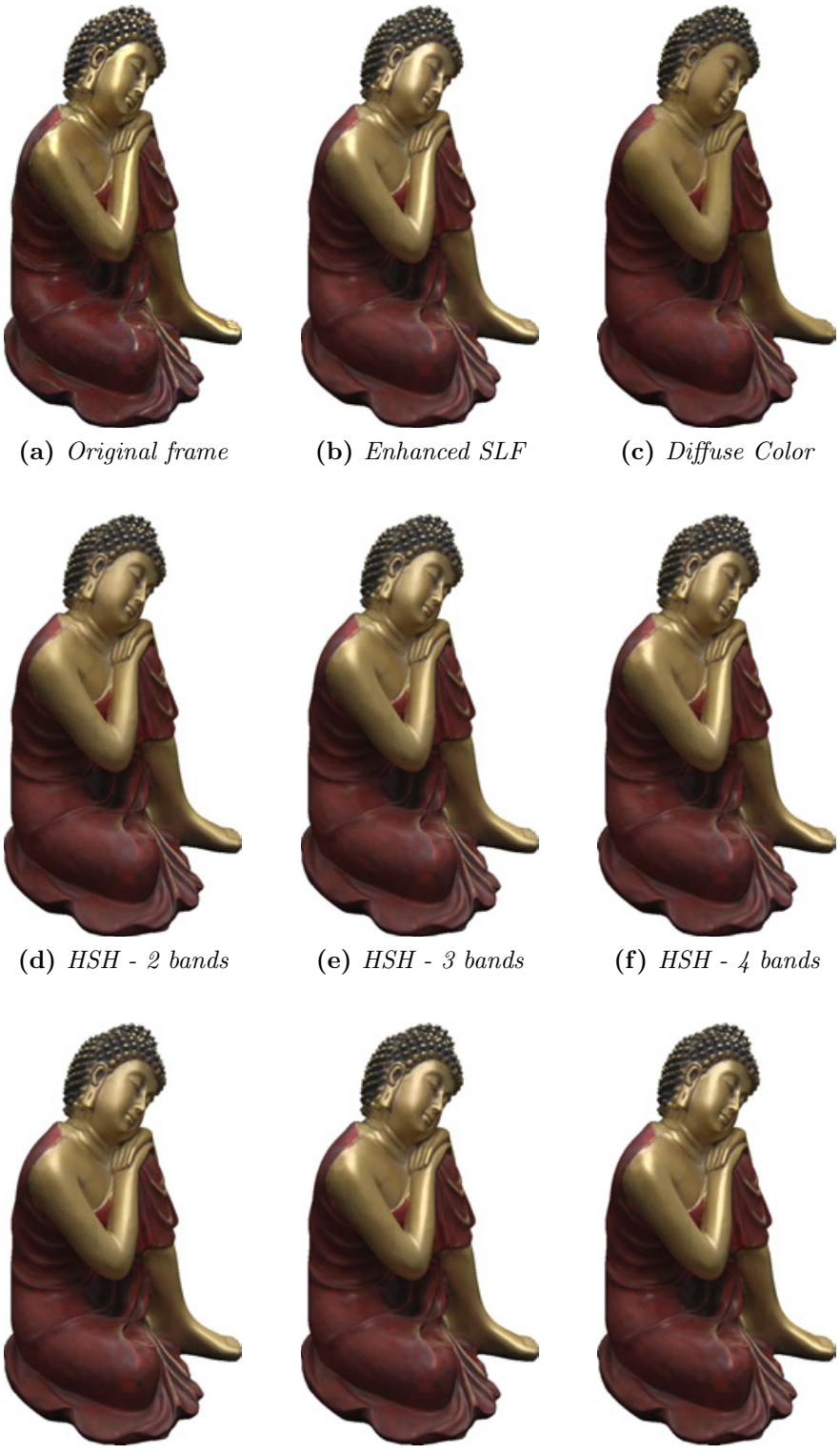
Figure 4.7: *DWARF results*

(a) *Original frame* **(b)** *Enhanced SLF* **(c)** *Diffuse Color*

(d) *HSH - 2 bands* **(e)** *HSH - 3 bands* **(f)** *HSH - 4 bands*

(g) *SH - 3 bands* **(h)** *SH - 4 bands* **(i)** *SH - 5 bands*

Figure 4.8: *GNOME results*



(a) *Original frame* **(b)** *Enhanced SLF* **(c)** *Diffuse Color*

(d) *HSH - 2 bands* **(e)** *HSH - 3 bands* **(f)** *HSH - 4 bands*

(g) *SH - 3 bands* **(h)** *SH - 4 bands* **(i)** *SH - 5 bands*

Figure 4.9: *SLEEPING BUDDHA results*

component:

$$C(u,v,s,t) = D^{(u,v)} + I_s \sum_{i=0}^{n} x_i^{(u,v)} h_i(s,t) \tag{4.11}$$

In this way, the algorithm can cope with the fact that the narrow specularity peaks are sometimes lowered by the band-limited reconstruction. The rendering results with this new parameter I_s are shown in the Figures 4.7b, 4.8b and 4.9b, using the SLF estimated with 5 bands of Spherical Harmonics (Figures 4.7i, 4.8i and 4.9i). The values assign to the parameter I_s are 1.6 for the GNOME and the DWARF and 1.2 for the SLEPPING BUDDHA. The visual differences between the rendering and the original frame are further reduced. This observation is confirmed from the estimation of the MSE and SSIM metrics between the two images, reported in the last row of the Table 4.2.

Figure 4.10 shows some additional renderings of the objects from different points of view that have not been covered by the video sequences. The rendering results are visually correct without artifacts.

Figure 4.10: *Additional rendering results from points of view not acquired by the input video sequences: (Left) DWARF; (Center) GNOME; (Right) SLEEPING BUDDHA.*

Chapter 5

Spatially Varying BRDF Statistical Estimation

This chapter presents a statistical method for the estimation of the Spatially Varying Bidirectional Reflectance Distribution Function (SVBRDF) of an object with complex geometry, starting from video sequences acquired with fixed but general lighting conditions. The aim of this work is to define a method that simplifies the acquisition phase of the object surface appearance and allows the reconstruction of an approximated SVBRDF. The final output is suitable to be used with a 3D model of the object to obtain accurate and photo-realistic renderings. The method is composed by three steps: the approximation of the environment map of the acquisition scene, using the same object as a probe; the estimation of the diffuse color of the object; the estimation of the specular components of the main object materials, using a Phong model. All the steps are based on statistical analysis of the color samples projected by the video sequences on the surface of the object. Although the method presents some limitations, the trade-off between the easy of acquisition and the obtained results make it a good choice for practical applications.

5.1 SVBRDF from videos

The acquisition of the surface appearance, together with the reconstruction of the 3D geometry, is a fundamental step for the photo-realistic rendering of real objects. Although several methods have been proposed for the acquisition of the SVBRDF of objects with a complex geometry, they present some limitations: lack of flexibility, due to solutions developed to solve specific problems that impose some constrains on the type of the material to acquire, for example the acquisition of the reflectance of the human face; the use of specialized instruments, like dome or special setup for the camera and the light source, characterized by complex mechanical, optical and electronic configurations that make difficult to extend them on objects of arbitrary shape and size; the high cost for the realization of these specialized acquisition

devices; the need of highly controlled acquisition environments, such as a dark room, which are difficult to reproduce in the case the environment in which the object is placed cannot be modified and the object cannot be moved (museums, outdoor locations, etc.).

This chapter introduces a new method for an approximate SVBRDF reconstruction that simplifies the acquisition setup for two main aspects: it uses a single and standard camcorder to acquire video streams; it performs the acquisition in a general lighting environment without the need of special lighting conditions. The main advantages of the method in the acquisition step are the use of cheap and widely diffuse devices, like a common video camera, and the lowering of the expertise of the operator that does not worry about the creation of complex lab lighting condition. The method takes in input video sequences and estimates the surface appearance of the object of interest. The videos are acquired with fixed and unknown lighting conditions, moving a Low Dynamic Range (LDR) video camera around the object. Starting from the alignment of the video streams on a previous acquired 3D geometry of the object, the target is to extract as much as possible information from the video frames to recover both the illumination conditions and the SVBRDF. The estimation is based on statistical analysis of the redundant video color data that is projected on the geometry.

The main contribution is a new method to estimate the SVBRDF of the object with the following features:

- an easy acquisition step for fixed but general lighting conditions;
- the approximation of the acquisition environment map with a weighted accumulation approach, using the object itself as a probe and taking advantage of a temporal coherent multi-view acquisition;
- a statistical method for the estimation of SVBRDF from the video color data projected on the geometry, taking advantage of the temporal coherence and the data redundancy of a video sequence and using a straightforward solution without time-consuming iterative and optimization procedures.

5.2 System Overview

The input of the proposed algorithm is a triangulated mesh of a real-world object with an associated parameterization and some LDR videos, which are acquired moving the camera around the object. The method assumes that the videos have been acquired with the camera in manual mode, with fixed exposure and fixed white balance. The white balance can be set by the user with the automatic procedure available on the camera using a reference white object, in order to correct the color of the lights. The color video output is assumed to be in sRGB color space and the linear response of the camera is obtained by conversion of the sRGB output in

the CIE LAB color space, a perceptually uniform space where an uniform change of the color value correspond to an uniform change of the perceived color. This feature allows the approximation of the perceptual difference between two colors by taking the simple Euclidean distance between them. The algorithm can work with complex illumination environments composed by different kind of light sources (e.g. point/area light, positional/directional light), which have a fixed position and a nearly constant intensity during the acquisition.

To model the surface appearance, the algorithm use a simple spatially varying Phong model, where each texel (x, y) of the mesh parameterization is characterized by the sum of a Lambertian diffuse component $D_{x,y}$ and a specular lobe $S_{x,y}$:

$$\rho_{x,y}(\vec{v}, \vec{l}) = D_{x,y}(\vec{l}) + S_{x,y}(\vec{v}, \vec{l}) = \rho_d(\vec{l} \cdot \vec{n}) + \rho_s(\vec{l} \cdot \vec{r})^{\alpha} \tag{5.1}$$

where $\vec{v}$ is the view direction, $\vec{l}$ is the light direction, $\vec{n}$ is the surface normal, $\vec{r}$ is the mirror reflected vector of the view direction with respect to $\vec{n}$, ρ_d is the color of diffuse component, ρ_s is the color of the specular highlight and α is the shininess of the highlight. The output of the algorithm is the diffuse texture map, the clustering of the texels in a set of basis materials and the specular parameters ρ_s and α for each basis material. The algorithm is composed by the following steps:

1. Registration of the videos over the mesh (Section 5.3)
2. Estimation of the environment map (Section 5.4)
3. Estimation of the diffuse color per texel (Section 5.5)
4. Estimation of the specular parameters per material (Section 5.6)

All the computations in the steps 2, 3 and 4 are done in GPU. Figure 5.1 shows an overview of the system.

In order to make more tractable the problem from a statistical point of view, the algorithm makes three assumptions on the lighting environment and the material composition of the object:

- all the lights have almost the same color, not necessarily white because it is generally corrected by the white balance procedure available on the camera;
- the lights have almost the same intensity in terms of order of magnitude, because the estimated diffuse color includes the light intensity;
- the possibility to cluster the object's materials with a user-assisted method starting from an automatic segmentation of the diffuse color (the proposed method uses a generic growing regions procedure, but any other approach can be applied to solve more challenging cases).

Even if the first two assumptions are very strong, these conditions are quite common in the real world acquisition environments, like building rooms, museums or outdoor locations.

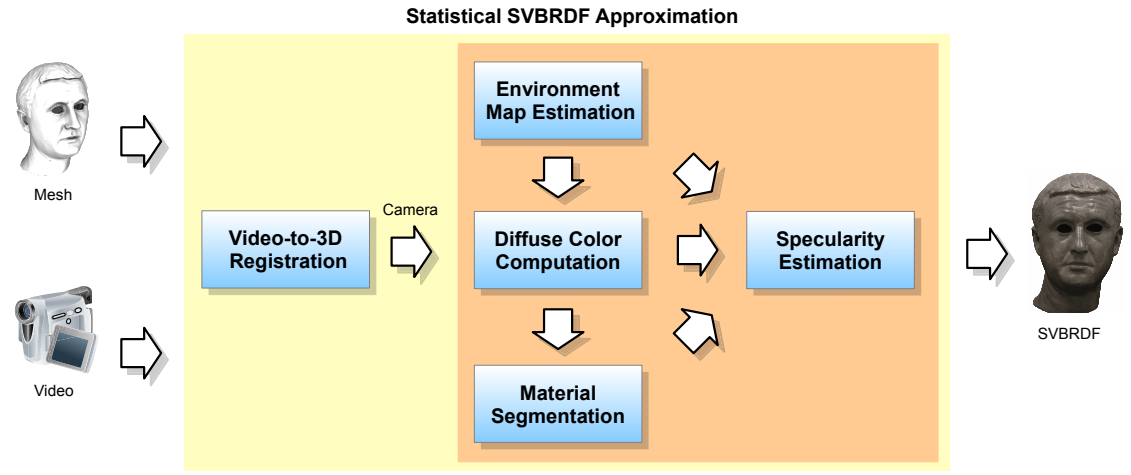


Figure 5.1: *Statistical SVBRDF approximation: algorithm overview.*

5.2.1 Visibility Approximation

In all the steps, the algorithm needs a fast way to compute the visibility along a given direction for each object surface point, in order to manage the effects of self-occlusion and self-shadowing. For this reason the algorithm precomputes a spherical harmonics approximation of the visibility function $V : \vec{\omega} \quad \Omega \quad \{0,1\}$. In the specific, it approximates the function for each vertex i with $\tilde{V}_i$, using the first 6 bands of spherical harmonics $Y^{(l)(m)}$ (36 coefficients $k_i^{(l)(m)}$):

$$\tilde{V}_i : \vec{\omega} \quad \Omega \quad [0,1] \tag{5.2}$$

$$\tilde{V}_i(\vec{\omega}) = \sum_{l=0}^{5} \sum_{m=-l}^{l} k_i^{(l)(m)} \, Y^{(l)(m)}(\vec{\omega})$$

The function $\tilde{V}_i$ returns the probability that there is visibility along each direction $\vec{\omega}$ around the vertex i. The visibility value $\tilde{V}_{x,y}(\vec{\omega})$ of each texel (x, y) is obtained by barycentric interpolation among the values of the triangle vertices. Even if the function $\tilde{V}$ is an approximation, the first 6 bands of spherical harmonics are sufficient to have a good reconstruction of the real visibility.

5.3 Video-to-3D Geometry Registration

The registration of the videos over the mesh is obtained using the algorithm presented in the Chapter 3. In order to obtain the best projection of the color samples on the geometry and then estimate a high quality approximation of the SVBDRF, the best possible video-to-geometry alignment is obtained by forcing the registration algorithm to execute the alignment procedure by Mutual Information, introduced in the Section 3.2.3, on each frame. The output of the registration algorithm is the

perspective camera projection matrix of each video frame. These matrices allow the recovering of the set of color samples $C_{x,y} = \{I^{(j)}_{x,y} \in RGB\}$ projected by each frame j over the texel (x, y). For each sample, the respective quality value $q^{(j)}_{x,y}$ is computed using the framework proposed in [23] and already presented in the Section 4.2.1. The quality is equal to the product of three measures normalized in the range $[0, 1]$ (Figure 5.2): the distance in image space from the nearest depth discontinuity (Figure 5.2a), to penalize the wrong color samples due to small misalignment; the depth of the texel in camera space (Figure 5.2b), to assign a higher quality at the pixels that are sampled by closer views; the dot product between the view direction and the surface normal (Figure 5.2c), to give a higher quality at the samples viewed by a less orthogonal directions.

(a) **(b)** **(c)** **(d)**

Figure 5.2: *Per pixel quality function: (a) map of the border distance from the depth discontinuities; (b) depth map; (c) dot product between the normal and the view direction; (d) final quality.*

5.4 Environment Map Reconstruction

As already done in the Chapter 4, for the estimation of the environment map, the algorithm uses an accumulation approach based on a statistic separation of the color samples that could have a specular behavior. If the input video sequences were acquired in ideal working condition (a single controlled light and a camera with linear response and without noise), the algorithm could assign as diffuse color the projected sample with minimum luminance. Since this is not the case and the residual and registration error generate wrong projected color samples, especially near the color boundaries, the algorithm is based on the computation of a luminance threshold to separate, for each texel, the specular and the diffuse samples. The first ones are used in the estimation of the environment, the last ones in the computation of the diffuse color.

A good and usual choice in the multi-view dataset for the separation between

specular and diffuse components is the computation of the median of the luminance. The main drawback in this computation is the linear memory occupancy, with respect to the number of samples projected on the texel, needed for the sorting of color samples, making it less suitable for a fast computation in GPU. For this reason the algorithm computes for each texel (x, y) a luminance threshold $t_{x,y}$ that is an upper bound of the median (see Appendix A.1 for more details) and is equal to the sum of mean and absolute deviation of the luminance of the color samples $C_{x,y}$ projected on the texel. More in details:

$$t_{x,y} = \mu_{x,y} + \bar{\sigma}_{x,y} \tag{5.3}$$

where $\mu_{x,y}$ and $\bar{\sigma}_{x,y}$ are the mean and the average absolute deviation:

$$\mu_{x,y} = \frac{\sum_{I^{(j)} \in C_{x,y}} Lum(I^{(j)})}{\#(C_{x,y})} \tag{5.4}$$

$$\bar{\sigma}_{x,y} = \frac{\sum_{I^{(j)} \in C_{x,y}} \left| Lum(I^{(j)}) - \mu_{x,y} \right|}{\#(C_{x,y})} \tag{5.5}$$

where $Lum(I^{(j)})$ returns the luminance of the color sample $I^{(j)}$ in the CIE LAB color space. To be more robust in the computation of this threshold, the algorithm discards all the color samples with a quality lower than 0.1. All the color samples with a luminance above the threshold $t_{x,y}$ are used in the estimation of the environment map.

For each element in $A_{x,y} = \{I^{(j)} \in C_{x,y} \mid Lum(I^{(j)}) > t_{x,y}\}$, the algorithm computes the specular mirror direction $\vec{r}$ of the view vector $\vec{v}_{x,y}^{(j)}$ with respect to the texel surface normal $\vec{n}_{x,y}$ and then it accumulates an amount $a_{x,y}$ in the environment map along this direction. The value $a_{x,y}$ is equal to the difference of the sample luminance from the threshold $t_{x,y}$ multiplied by the sample quality $q^{(j)}$ and the visibility along the direction $\vec{r}$.

$$a_{x,y} = (Lum(I^{(j)}) - t_{x,y})\ \tilde{V}_{x,y}(\vec{r})\ q^{(j)} \tag{5.6}$$

The computation of the value $a_{x,y}$ and the availability of a multi-view acquisition give more robustness with respect to possible geometric and registration errors. Then, the map is normalized with the distribution of all color samples along the specular mirror direction in the environment. This distribution is computed as the total number of the color samples in $C_{x,y}$ that project in each specific pixel of the environment map. This normalization gives more robustness with respect to the camera movement, in term of temporal and spatial density of the acquisition. In this construction, the algorithm assumes that the specularity is located where the object behaves like a mirror and, even if the mirror specular direction does not

entirely match with the real specular direction, it is usually very close. Finally, the environment map is normalized in the range $[0, 1]$. The result is a probability map depicting the position of the lights in the environment.

In this step, the object itself is used as a probe without the need to put any other usual light probe objects (e.g. spheres and parabolic mirror) in the scene. Although the object does not have optimal features in term of angle coverage and uniform sampling of the space, the results obtained show that this approach is sufficient to describe the lighting environment with enough detail and precision (Figure 5.6).

5.5 Diffuse Color Estimation

The estimated environment map allows the computation of a new weight for each color sample, the specular weight, which represents the probability that the sample has a specular behavior. This weight is useful for the estimation of the diffuse color. It is computed by directional sampling of the environment map in a cone of directions along the specular mirror direction $\vec{r}$, followed by a normalization that depends on the luminance value of the samples. The method uses a cone of direction to take into account two problems: materials with a partially off-specular behavior; the geometric and registration errors.

Given a set of 16 directions $B = \{\vec{\omega} \in \Omega \mid \vec{\omega} \cdot \vec{r} < \gamma\}$ distributed around $\vec{r}$ with a stratified disk sampling strategy, the specular weight $s_{x,y}^{(j)}$ for each color sample is:

$$s_{x,y}^{(j)} = \frac{1}{w} \, (b_{x,y}^{(j)})^{w^2} \tag{5.7}$$

with:

$$b_{x,y}^{(j)} = \frac{\sum_{\vec{\omega} \in B} g(\vec{\omega} \cdot \vec{r}) \; \tilde{V}_{x,y}(\vec{\omega}) \; EnvMap(\vec{\omega})}{\sum_{\vec{\omega} \in B} g(\vec{\omega} \cdot \vec{r}) \; \tilde{V}_{x,y}(\vec{\omega})} \tag{5.8}$$

$$w = \frac{t_{x,y} - min_{x,y}}{Lum(I_{x,y}^{(j)}) - min_{x,y}} \tag{5.9}$$

where the function $EnvMap(\vec{\omega})$ returns the value of the environment map along the direction $\vec{\omega}$, g is a Gaussian function with standard deviation $\sigma = \gamma = \cos 5°$ to weight the contribution of each direction with respect to the distance from the direction $\vec{r}$, $\tilde{V}_{x,y}$ computes the visibility along $\vec{\omega}$, $min_{x,y}$ is the minimum luminance value projected on the texel, $t_{x,y}$ is the threshold computed in the Equation 5.3 and $Lum(I_{x,y}^{(j)})$ is the luminance of the current sample. The term $b_{x,y}^{(j)}$ is the weighed average of the values gathered from the environment map along the directions in B and it represents the probability that there is a light source around the specular mirror direction. The term $\frac{1}{w}$ and the exponent w^2 are two texel-aware correction factors, which depend on the distribution of all color samples around the threshold

$t_{x,y}$ and the position of current sample in this distribution. These terms provide more robustness with respect to the real reflectance behavior of the sample, detecting the false positive cases (samples with a diffuse behavior that are geometrically aligned with an area of the environment map associated to a strong light source).

For the computation of the diffuse color, the algorithm adapts the same approach used in the estimation of the environment map. For each texel it computes a threshold $d_{x,y} = \mu_{x,y} + \bar{\sigma}_{x,y}$, where $\mu_{x,y}$ and $\bar{\sigma}_{x,y}$ are the weighted mean and the weighted absolute deviation with the specular weights $s^{(j)}_{x,y}$:

$$\mu_{x,y} = \frac{\sum_{I^{(j)} \in C_{x,y}} (1.0 - s^{(j)})\ Lum(I^{(j)})}{\sum_{I^{(j)} \in C_{x,y}} (1.0 - s^{(j)})} \tag{5.10}$$

$$\bar{\sigma}_{x,y} = \frac{\sum_{I^{(j)} \in C_{x,y}} (1.0 - s^{(j)}) \left| Lum(I^{(j)}) - \mu_{x,y} \right|}{\sum_{I^{(j)} \in C_{x,y}} (1.0 - s^{(j)})} \tag{5.11}$$

The final diffuse color $D_{x,y}$ is equal to the weighted average of the color samples with a luminance lower than the threshold $d_{x,y}$:

$$D_{x,y} = \frac{\sum_{I^{(j)} \in A} q^{(j)} I^{(j)}}{\sum_{I^{(j)} \in A} q^{(j)}} \tag{5.12}$$

where $A_{x,y} = \{I^{(j)} \in C_{x,y} \mid Lum(I^{(j)}) < t_{x,y}\}$ and $q^{(j)}$ is the quality of the samples. The weighted average of the sample in $A_{x,y}$ is justified by the need to balance out the not ideal acquisition conditions (geometric and registration error, camera with noise and not linear response).

5.6 Specularity Estimation

The final step of the system is the estimation of the specular behavior of the object. Starting from the observation that the majority of the real objects have a specular component that is less spatially varying than the diffuse color, the algorithm assumes that the object can be segmented in a set of basis materials, each one with a different specular reflectance. For this purpose, it completes the following tasks:

1. Detection of the main light sources in the environment map;

2. Estimation of the specularity parameters per texel;

3. Clustering of the basis materials of the object based on the diffuse color;

4. Estimation of the specularity parameters per cluster.

The main idea under this step is that the slope of a temporal coherent luminance peak can be used to estimate the two specular parameters ρ_s and α (see Equation 5.1 in Section 5.2) for each basis material (Figure 5.3).

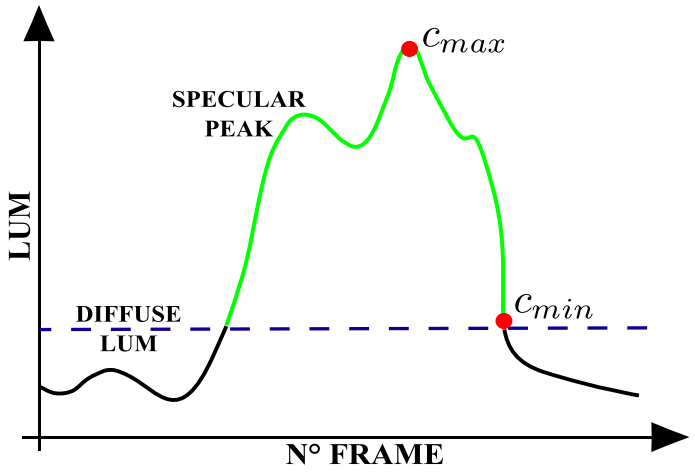


Figure 5.3: *Temporal trend of the texel luminance. The figure highlights the luminance of the diffuse color, the temporal coherent luminance peak (green line), and the color samples c_{min} and c_{max} used in the estimation of the specular parameters per texel.*

The first task is the estimation of the position and shape of the main light sources in the scene by clustering. The algorithm starts by approximating the environment map with a set L of 4096 directional lights, using a median cut algorithm [43]. The algorithm distributes a set of 4096 directional lights in the map with a density that depends on the pixel intensity. The higher the pixel intensity is, the denser is the local distribution of the directional lights. Each directional light approximates a small area of the environment. Then, it discards all the lights that cover an area of the environment greater than a tile of a sphere subdivided in 4096 uniform area regions ($4\pi/4096$). Finally, it clusters all the lights that create a connected component with their areas in the map. Each cluster represents a light source in the scene. For each cluster, it computes the centroid and the local distribution around the centroid to have a rough estimation of two important features of the illuminant (Figure 5.4): the position $\vec{l}$ and the shape. The centroid is computed using the image moment of the area on the map covered by the cluster. For the local distribution, it computes the covariance matrix Σ_l of the directional lights in the cluster with respect to the centroid, using an isometric mapping of the environment sphere on the tangent plane at the centroid. The covariance matrix is used to approximate the distance of a direction from a light source using the Mahalanobis distance. In

this clustering process, the algorithm assumes that all the lights have the same intensity. This assumption creates problems only when the environment has a very large dynamic range that does not allow observing all the main reflectance features of the object appearance.

The second step is the estimation of the specular parameters ρ_s and α for each texel. These parameters are computed using only the luminance channel data, assuming that all the lights are white. In the specific, for each texel the algorithms detects the temporal coherent luminance peaks, which is a consecutive sequence of color samples in the video that have a luminance greater than the luminance of the diffuse color $D_{x,y}$. Then, for each peak it selects the two samples c_{min} and c_{max} with the minimum and the maximum luminance difference from the diffuse color (Figure 5.3). In this process, it discards all the samples with a luminance value higher than 98 (out of 100) to be more robust with respect to the saturation of camera CCD sensor, where the camera response is highly not linear. Following, it assigns to each luminance peak the main light source $\vec{l} \in L$ that should produce it, which is the light source from which the reflected view direction $\vec{r}_{max}$ of the maximum luminance sample c_{max} has the lowest Mahalanobis distance. The Mahalanobis distance allows the algorithm to take account of the illuminant shape. With this data, the algorithm can compute the parameters ρ_s and α solving the following system of equations (see Appendix A.2 for more details):

$$\begin{cases} \rho_s(\vec{l}\cdot\vec{p}_{max})^{\alpha} &= Lum(c_{max}) - Lum(D_{x,y})) \\ \rho_s(\vec{l}\cdot\vec{p}_{min})^{\alpha} &= Lum(c_{min}) - Lum(D_{x,y})) \end{cases} \tag{5.13}$$

where $\vec{p}_{max}$ and $\vec{p}_{min}$ are the directions that minimize the dot product with the light direction and have the same Mahalanobis distance of the mirror reflected view directions $\vec{r}_{max}$ and $\vec{r}_{min}$ (Figure 5.4):

$$\begin{aligned} \vec{p}_{max} &= \arg\min_{\vec{v}\in A}(\vec{l}\cdot\vec{v}) \\ \vec{p}_{min} &= \arg\min_{\vec{v}\in B}(\vec{l}\cdot\vec{v}) \end{aligned} \tag{5.14}$$

with:

$$\begin{aligned} A &= \{\vec{\omega}\in\Omega \mid d_{Mahal}(\vec{\omega},\Sigma_l) = d_{Mahal}(\vec{r}_{max},\Sigma_l)\} \\ B &= \{\vec{\omega}\in\Omega \mid d_{Mahal}(\vec{\omega},\Sigma_l) = d_{Mahal}(\vec{r}_{min},\Sigma_l)\} \end{aligned} \tag{5.15}$$

In this process, the algorithm discards all the peaks where the Mahalanobis distance d_{Mahal} of the sample c_{min} is less than the Mahalanobis distance of the sample c_{max}, because it is an inconsistent case where the sample with a higher luminance is statistically farther from the mirror reflected light directions than the sample with the lower luminance. If a texel has more than one peak, the algorithm chooses the one with the minimum Mahalanobis distance from its main light, which is the peak that minimizes the dot product between the view direction of the sample c_{max} and the mirror reflected direction of its main light.

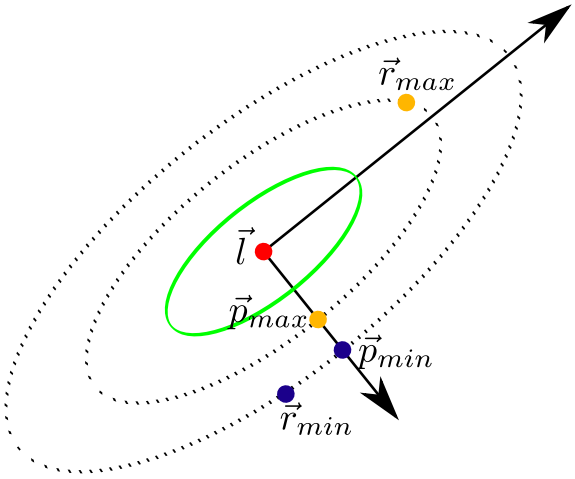


Figure 5.4: *Characterization of a light source with the direction $\vec{l}$ and the local distribution defined from the covariance matrix Σ_l (green ellipse). The characterization allows the projection of the mirror reflected view vectors $\vec{r}_{max}$ and $\vec{r}_{min}$ on the direction of minimum variance ($\vec{p}_{max}$ and $\vec{p}_{min}$) defined by the covariance matrix Σ_l.*

Due to the fixed lighting conditions and the limited view directions covered with the video sequences, it is possible that some texels do not have enough data to estimate the specular parameters. To overcome this problem and to obtain a more consistent and robust estimation of the specular reflectance, the algorithm employs a partial user-assisted method to create clusters of points with the same specular behavior. This clustering is based on the diffuse color of the texels. The main idea is to create a new cluster for each area of the object with a uniform perceived diffuse color that has a singular specular behavior.

The clustering starts with an automatic step that creates an initial material segmentation. It distributes a set of seed points in the areas with the most uniform diffuse color and then applies a growing region for each of these seeds. A new point is added to the cluster if the distance of its color from the mean color of the cluster is below a threshold selected by the user. The distance between two colors is computed as euclidean distance in the CIE LAB color space. Starting from this initial segmentation, a simple interactive application allows the selection of a cluster, by picking a point on the object, and the execution of two different operations: to merge a set of clusters; to split a cluster along a stroke drawn by the user over the object. These operations solve two different challenging cases: over-segmentation of an area with uniform material; separation of adjacent areas with the same diffuse color and different specular behavior.

The final step is the estimation of the specular parameters for each cluster, using the data computed for the texels in the cluster. The most challenging parameter to estimate in this step is ρ_s due to several reasons: the clamped luminance signal returned by the LDR video camera; the highly spatially varying nature of the dif-

fuse color of the texel in the cluster; a bad clustering of the texel especially on the boundary among the clusters. To overcome these problems, the algorithm employs a statistical analysis of the ρ_s values of the texels for each cluster independently. It computes the probability density function of the ρ_s values in the cluster using a kernel density estimation method [167]. The result is a multi-modal probability density function where the algorithm looks for the influence area of the higher mode, defined by its statistical bell. Finally, it computes the average of the specular parameters ρ_s and α for all the texels in the cluster that have a ρ_s value inside this statistical bell.

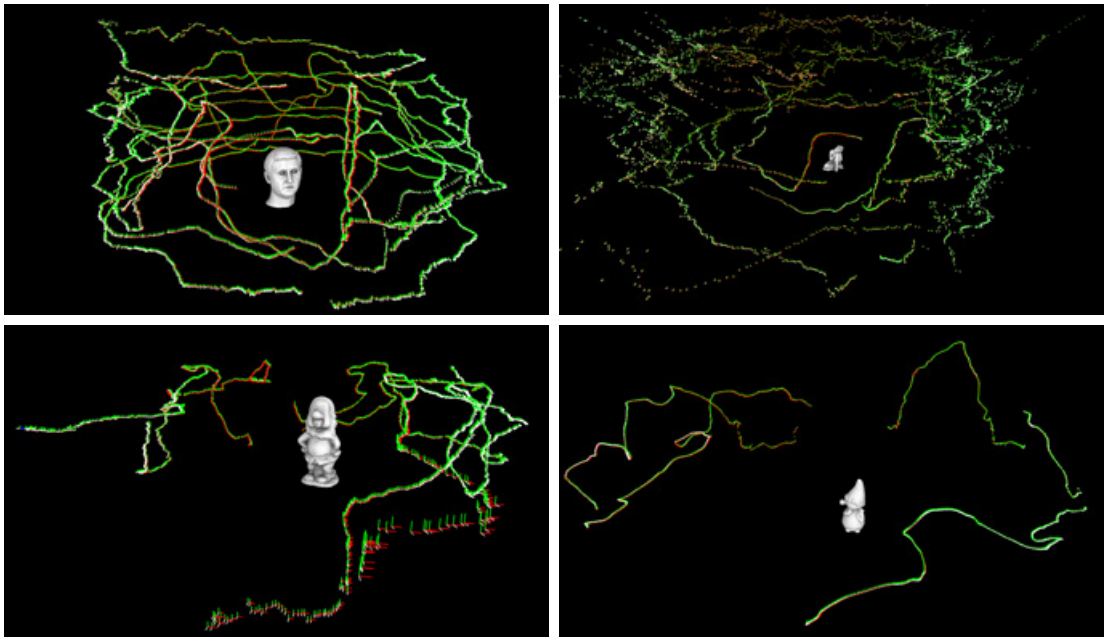


Figure 5.5: *Test cases camera path obtained by the registration algorithm: (Top-Left) HEAD; (Top-Right) SLEEPING BUDDHA; (Bottom-Left) DWARF; (Bottom-Right) GNOME.*

5.7 Results

The algorithm has been tested with four different objects of different materials and different reflectance behaviors:

- the DWARF, a terracotta statue (30cm tall) that presents different types of specularity, in size and intensity: sharper and with a high-medium intensity on the dress; wider on the face; almost completely absent on the beard;

- the GNOME, a ceramic statue (15cm tall) that has very sharp and high specularity on the hat and a near diffuse behavior on the body;

- the HEAD, a bronze copy of the head of the Arringatore, an Etruscan statue from the National Archeological Museum in Florence (30cm tall);

- the SLEEPING BUDDHA, an acrylic resin Buddha (10cm tall) with different types of coatings (a gold paint on the body, a reddish specular paint on the dress and a diffuse black paint on the hair).

The video sequences for the HEAD and the SLEEPING BUDDHA present a more uniform sampling of the view direction, while the sequences for the DWARF and the GNOME are more sparse (Figure 5.5).

The 3D models were acquired by 3D scanning, using a Konica Minolta VI-910 laser scanner, and then simplified to obtain a medium resolution model for the computation of the texture parameterization [154]. In this process, the normal map of the high resolution mesh is saved in order to preserve the details lost during the simplification.

The videos were acquired with a full HD video camera that is set at the highest quality to reduce the compression artifacts. To make the estimation of the appearance more robust, the camera is set in manual mode with a fixed white balance, defined with the automatic procedure available on the camera using a reference white object, and fixed exposure.

Two different lighting environments are used for the acquisition. The two environments are characterized by different type of lights: the first scenario (Figure 5.6a) is characterized by three near-circular halogen floodlights; the second scenario (Figure 5.6d) is characterized by six fluorescent tubes. The DWARF and the GNOME were acquired in the first scenario, while the HEAD and the SLEEPING BUDDHA were acquired in the second scenario.

5.7.1 Environment Map

Figure 5.6 shows a comparison between the real environment maps, taken during the acquisition with a metal reflective sphere, and the reconstruction obtained from the four test cases data using the procedure described in the Section 5.4. In general, the method does not reconstruct the entire environment map, but only the main light sources that produce a specularity on the object. More in details, in the first scenario it recovers the three main light sources, both in shape and position. In the second scenario, the method detects only the two lights that were over the objects, neglecting the other lights that are too far to produce highlights that can be recorded in the video during the acquisition.

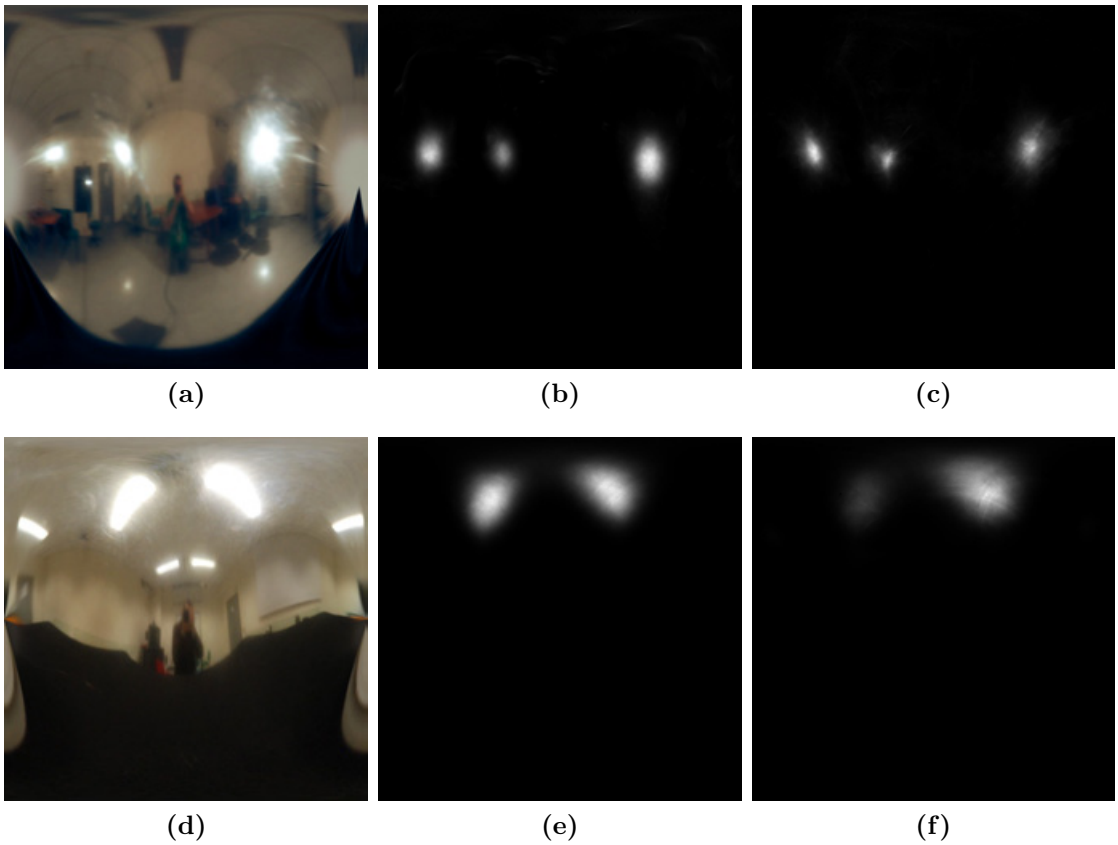


Figure 5.6: *Environment maps results: (a) first real environment scenario; (b) first environment scenario reconstructed from the DWARF's videos; (c) first environment scenario reconstructed from the GNOME's videos; (d) second real environment scenario; (e) second environment scenario reconstructed from the HEAD's videos; (f) second environment scenario reconstructed from the SLEEPING BUDDHA's videos.*

5.7.2 SVBRDF Appearance Approximation

Figure 5.7 shows a visual comparison between the rendering of the object with the obtained SVBRDF reconstruction and a frame of the video used in the estimation process. From left to right there are the clustering of the basis materials, the diffuse component, the specular component (normalized in the range $[0, 1]$ to improve its visualization), the final rendering and the original video frame without background. The renderings are generated using the environment maps estimated in Section 5.4, approximated with a set of 256 directional lights obtained through a median cut algorithm [43]. The final results show that the obtained SVBRDF reconstruction is able to reproduce the main features of the object appearance maintaining the relative differences among the several materials of the object. To remark that this reconstruction is not an absolute SVBRDF measurement, but a simple and fast

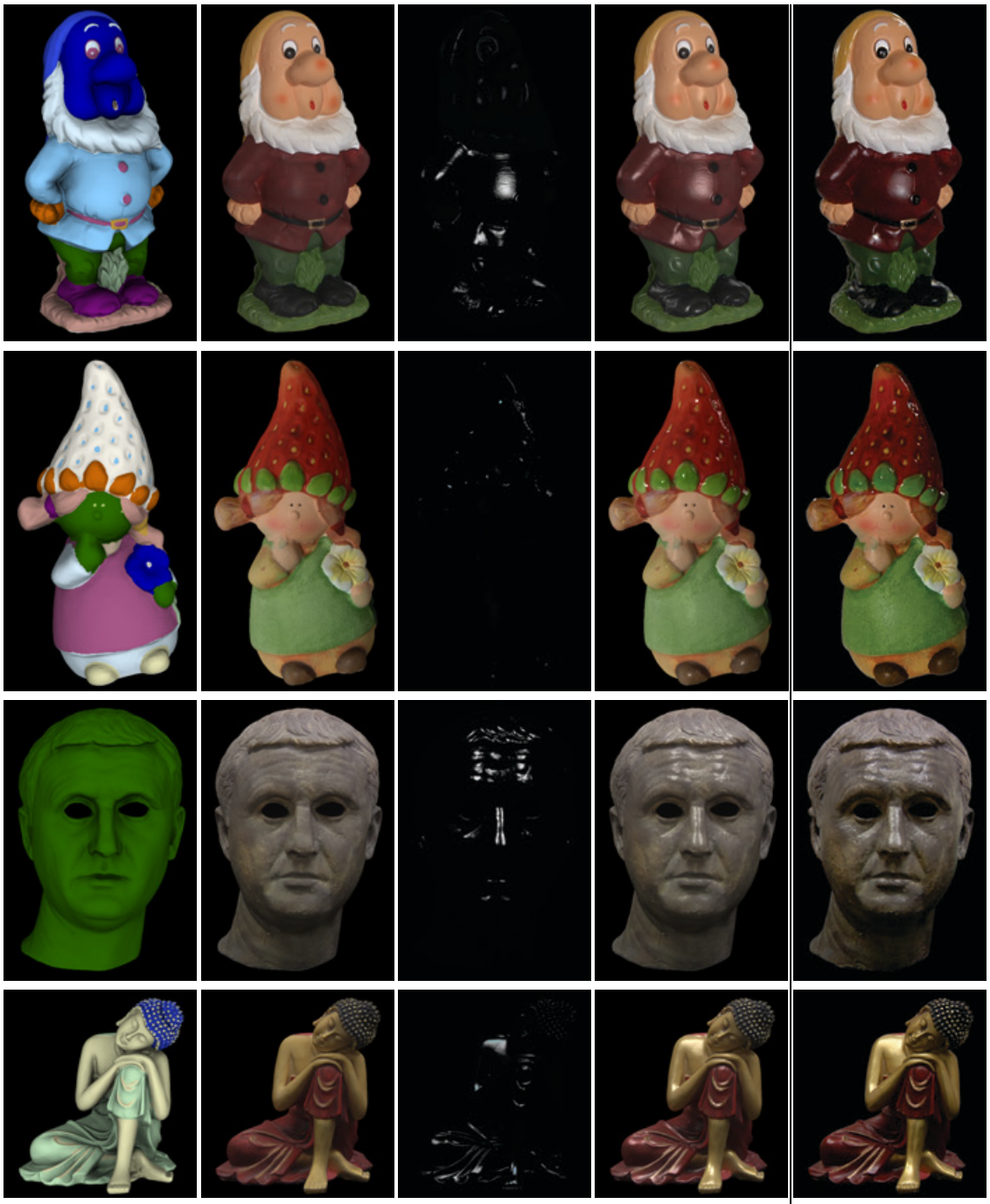


Figure 5.7: *SVBRDF rendering results. From top to bottom: DWARF; GNOME; HEAD; SLEPPING BUDDHA.*

approximation to allow a photo-realistic rendering of the object. For this purpose, the algorithm assumes that the video sequences have been acquired in a lighting environment with a good global illumination on the whole object that allows the observation of the main features of object appearance. This is obtained with limited input data in term of light and view sampling quality, which can be acquired in a short time in an uncontrolled environment. These conditions are usually not taken into account in the context of most of the material acquisition methods proposed in literature. Additional comparative renderings from different point of view are shown in the Figures 5.12 and 5.13.

Figure 5.11 shows a comparison between a photo of the object and the rendering with the SVBRDF reconstruction. The photos were acquired in a different lighting environment (a single halogen floodlight) with respect to the acquisition one, shot with the same video camera. The renderings show a good approximation of the object appearance, especially for the shape and position of the specularities, while small diffuse color shifts are due to the white balance procedure of the camera in the correction of the color of the light.

Figures 5.8, 5.9 and 5.10 display the rendering of the SLEPPING BUDDHA in three different HDR environment maps using a Monte Carlo ray tracer. The three selected environment maps are characterized by very different lighting conditions. The results are coherent and visual appealing, proving the good quality of the obtained SVBRDF approximation.

Due to the input data and the Phong model, the method presents some limitations: the reconstruction of a shaded diffuse color, due to the limit of the data acquired with fixed lighting conditions; a blur effect on the very small and sharp specularity, for example on the GNOME's hat, due to small misalignment in the registration step; the different appearance or the loss of some highlights, for example over the DWARF's hat, due to the lack of the surface meso-structure, to imprecise normals and to Fresnel effect at grazing angle that the Phong model is not able to reproduce; the final appearance is not independent from the lighting intensity of the acquisition environment. Anyway, even if there are some small visible differences to ground truth, the results do not present major artifacts and complex multi-material objects can be taken into account. The rendering results are effective and realistic making the method useful for practical applications that need a simple way to acquire and reproduce the appearance of a real object in real-time and in a photo-realistic manner.

5.7.3 Performance

Table 5.1 shows some data about the test cases and the processing time. For each case, it shows the total number of frames, the size of the 3D model, the time for the video-to-geometry registration, the time to reconstruct the environment map, the time to estimate the diffuse color and the specular parameters, and the number of basis specular materials. The experiments were performed on a PC with an Intel

	Frames	Geometry (triangles)	Registration (mm:ss)	Env. Map (mm:ss)	Diffuse (mm:ss)	Specular (mm:ss)	Clusters
DWARF	3382	200k	113:00	8:11	8:05	3:05	12
GNOME	2092	135k	73:00	5:08	5:03	1:57	13
HEAD	8386	250k	265:00	25:24	25:15	8:32	1
SLEEPING BUDDHA	7240	205k	251:00	18:57	18:49	6:21	3

Table 5.1: *SVBRDF approximation test cases data.*

Core i7 950 with 12GB of RAM and a NVIDIA GTX580 1536MB. The analysis of the table shows that most of the processing time stands in the registration step, while the rest of the algorithm is usually completed within minutes.

Figure 5.8: *HDR environment map rendering of the SLEEPING BUDDHA test case: Uffizi gallery.*

Figure 5.9: *HDR environment map rendering of the SLEEPING BUDDHA test case: Dining room of the Ennis-Brown House.*

Figure 5.10: *HDR environment map rendering of the SLEEPING BUDDHA test case: Pisa courtyard nearing sunset.*

Figure 5.11: *Comparison between an image of the object acquired in a different lighting environments (Left) and the rendering with our SVBRDF approximation (Right): (Top) HEAD; (Bottom) DWARF.*

Figure 5.12: *HEAD and SLEEPING BUDDHA results: (Left) rendering; (Right) original frame.*

Figure 5.13: *DWARF and GNOME results: (Left) rendering; (Right) original frame.*

Chapter 6

RTI: Shading Enhancement and Web Visualization

The Reflectance Transformation Imaging techniques are widely used for the documentation and the virtual analysis of CH artworks. Their success is due to several reasons: use of inexpensive and widely available hardware; simplicity of the acquisition process; scale well with the size of the artwork; support high sampling density and precision; require short processing time; produce photo-realistic rendering of challenging material like gold. This chapter presents two solutions proposed to address two open problems in the RTI field: shading enhancement and interactive web visualization for the dissemination. The first solution proposes a Multi-Lighting Detail Shading Enhancement that is based on the use of different light directions on different areas of the image. The second solution is an interactive web viewer for RTI images that takes advantage from recent web technologies, HTML5 and WebGL, to improve the dissemination of this new multimedia data and to support remote visual inspection of both scholar and ordinary public.

6.1 Reflectance Transformation Imaging

Thanks to their advantages over the 3D scanning techniques (inexpensive and widely available hardware, simple acquisition, scale well with the size of the artwork, high sampling density and precision, short processing time, photo-realistic rendering of challenging materials), RTI images are widely used as documentation tools and to support detailed visual analysis, giving a precious instrument to the Cultural Heritage specialists in the analysis and interpretation process.

The possibility to apply a set of operators to the RTI data to enhance the perception of the features of interest is an important improvement in the virtual analysis process of the artwork. Even if some simple methods have been already proposed (such as Diffuse Gain and Specular Enhancement in [127]), little attention has been devoted to the possibility to use the high amount of data compressed in this kind of

images. On the other hand the dissemination to the wide public of RTI technology, beyond the CH specialists, is still in an early status.

This chapter presents a new couple of shading enhancement operators, called Dynamic Multi-Lighting Enhancement and Static Multi-Lighting Enhancement. Using a different light direction for each area of the image, these operators improve the perception of details, features, and overall shape of the artwork. Finally, the chapter presents an interactive web viewer for RTI images that has been recently used in the realization of a museum kiosk for the presentation of the coin collection of the National Museum of San Matteo in Pisa.

6.2 Multi-Lighting Detail Enhancement

Multi-Lighting Detail Enhancement uses different light directions to illuminate the rendered image and increase the perception of surface details. When applied in the 2D RTI domain this visualization method works in real time and adaptively computes multiple light directions according to the portion of image under viewing and the light direction currently selected by the user. This enhancement method is able to reproduce what actually the archaeologists do locally by using a grazing light to better study and understand the artwork, preserving a good overall illumination. The advantage of this method is the possibility to create a virtual lighting environment with many lights, each one permitting to increase the local contrast of a small region of the object, which usually takes up some hundreds of pixel. This virtual lighting environment is not reproducible in the real world for several reasons: the number of lights; the very localized behavior of the light; the self-reflection effect of the object that usually decreases the sharpness of the image. The purpose of the method is to optimize an enhancement measure that maximizes the sharpness of the image and at the same time preserves the brightness. The result is a general improving of the perception of the shape and fine details in the RTI image.

6.2.1 Dynamic Multi-Lighting Enhancement

Typically, the visualization tools for RTI images allow the user to specify interactively the light direction. Zoom and pan operations are generally available to visualize and navigate high resolution images, usually together with a multi-resolution encoding computed with mip-mapping techniques, to speedup the rendering process.

Given the light direction l_0 that is selected by the user, and the section of the image currently rendered I^*, the Dynamic Multi-Lighting Detail Enhancement algorithm is composed by the following steps:

1. the image under viewing (I^*) is subdivided into $N \times M$ square tiles;

2. for each tile (T) l_0 is perturbed and the light direction (l') that maximizes the detail enhancement is chosen as the light direction for that tile;

3. the grid of $N \times M$ light directions obtained in this way are then made more uniform by applying a component-wise smoothing filter;

4. finally, per-tile light directions are converted into per-pixel light directions through bilinear interpolation.

The dimensions of the grid $M \times N$ depend on the tile size (in pixel) chosen by the user and are independent from the image resolution, because the algorithm exploits the mip-mapping system already used for the rendering of I^* according the current zoom factor. This strategy presents a performance advantage, because the number of tiles in any view remains more or less constant, and allows the re-computation of the lighting configuration at each zoom operation in order to reveals even more details.

The step 2 of the algorithm requires solving the following optimization problem for each tile of the grid:

$$l' = \arg\max_{l \in \mathcal{L}(l_0)} \mathcal{E}(T, l) \tag{6.1}$$

where $\mathcal{L}(l_0)$ is the set of perturbed light directions computed from l_0 and $\mathcal{E}$ is the enhancement measure defined in the Equation 6.2.

The set of lights $\mathcal{L}(l_0)$ is computed by perturbing l_0. The maximum perturbation can be chosen by the user up to 20°. Two different methods can be employed: anisotropic sampling and isotropic sampling. With the anisotropic sampling the set $\mathcal{L}(l_0)$ is generated such that the amount of perturbation decreases for the light directions that are almost tangent to the viewing plane (see Figure 6.1). This sampling strategy allows careful treatment of the case starting from a grazing light direction. In this condition, even small perturbations can result in high visual changes. Hence, in this case, the algorithm chooses to perturb the light direction with rotations along the viewing axis rather than with rotations perpendicular to it.

In practice, the anisotropic light perturbation takes advantage of the empirical knowledge that RTI images will probably represent something that has an interesting behavior under grazing light and therefore it employs a targeted optimization strategy that takes into account this critical configuration. On the other hand, the isotropic sampling spreads uniformly the light set in a cone of direction with axis l_0 that keeps the angular aperture independently from the direction of l_0. A comparison of the two methods is shown in Figure 6.3. In general, the anisotropic sampling preserves more of the dark-and-light parts of the image with respect to the standard rendering illuminated with grazing light, while avoiding visual artifacts. The isotropic sampling tends to illuminate the image more uniformly but it can potentially generate artifacts if abrupt lighting changes occur.

The enhancement measure $\mathcal{E}$ is defined as:

$$\mathcal{E}(T, l) = \alpha \mathcal{S}(T(l)) + (1 - \alpha)\mathcal{Y}(T(l)) \tag{6.2}$$

where $\mathcal{S}(.)$ is the sharpness of the tile T evaluated using a sharpness operators, $\mathcal{Y}(.)$ is a measure of the total brightness of the tile and α is a tuning parameter that

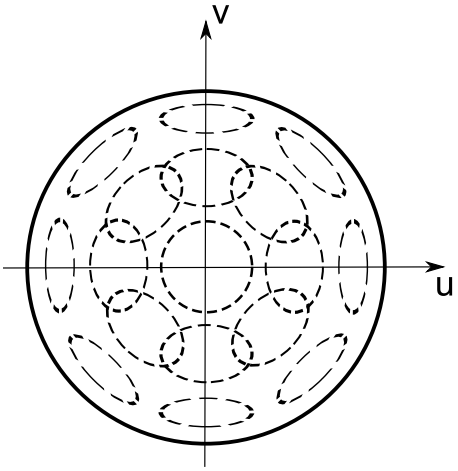


Figure 6.1: *Anisotropic sampling of the light direction. Each ellipse shows the direction sampled area assuming that l_0 is the ellipse center.*

controls the sharpness/brightness ratio. The brightness is evaluated by converting the RGB color component to YUV color space and by summing up the luminance for each pixel of the tile. The two measures are scaled according to the maximum of their values in order to have compatible values. The goal of the term $\mathcal{S}(.)$ is to increase the contrast of the image while the term $\mathcal{Y}(.)$ preservers the brightness of the image. The brightness term is very important because the sharpness maximization alone tends to exaggerate black-white contrast, making the final image too dark in some cases. Thanks to the brightness term the proposed algorithm is able to generate high-contrast images while preserving a good global illumination over the entire image (see Figure 6.6). The tested sharpness operators are:

$$M_1 = \iint \left|\frac{\partial I(x,y)}{\partial x}\right| + \left|\frac{\partial I(x,y)}{\partial y}\right| dx\,dy \tag{6.3}$$

$$M_2 = \iint \left|\frac{\partial I(x,y)}{\partial x}\right|^2 + \left|\frac{\partial I(x,y)}{\partial y}\right|^2 dx\,dy \tag{6.4}$$

$$M_3 = \iint \left(\nabla^2 I(x,y)\right)^2 dx\,dy \tag{6.5}$$

where M_1 is the L^1-norm of the image gradient, M_2 is the L^2-norm of the image gradient and M_3 is the energy of the image Laplacian. The image gradient is estimated with a Sobel filter, while the Laplacian is calculated using an image Laplace operator.

An example of the light configuration, computed as described, is shown in Figure 6.2.

After the lighting vectors are computed for each tile, a smoothing filter is applied. This filtering step is necessary since the light directions in adjacent tiles may differ considerably, producing visible artifacts. Even if the parameters of the filter can be

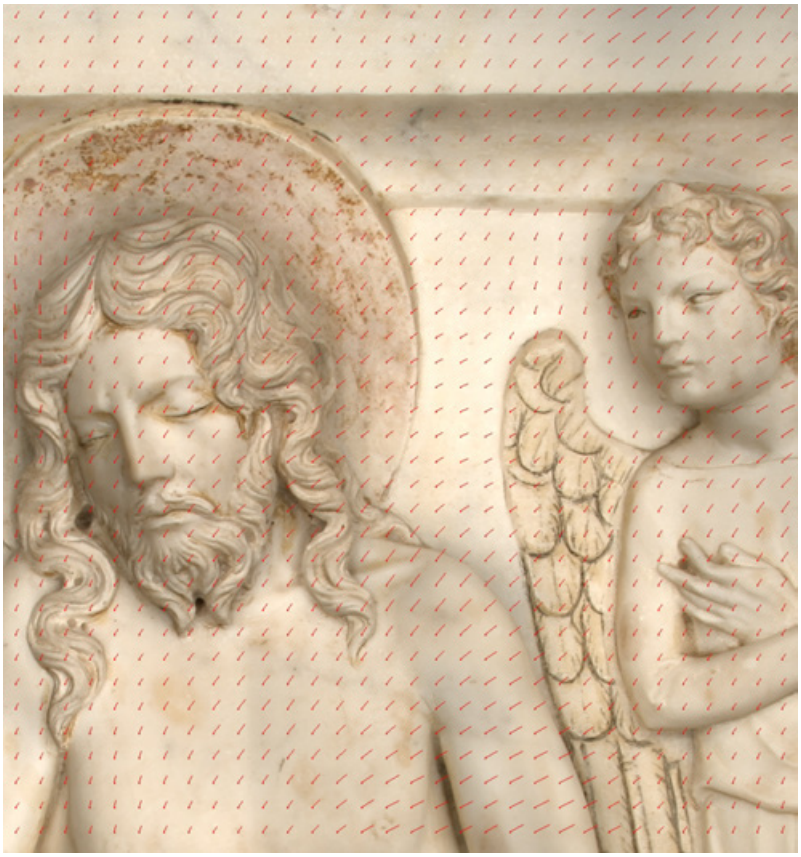


Figure 6.2: *Light configuration computed on-the-fly. Each vector represents the light direction used for that tile.*

chosen by the user, in many cases a box filter of size 3×3 tiles applied between 5 to 7 times produces a sufficient smoothing effect.

During the final rendering (step 4) a per-pixel light direction is used to avoid visual differences across the tiles. For each pixel the light direction is obtained by bilinear interpolation of the light directions of the four tiles adjacent to it.

One of the main advantages and innovative aspects of this enhancement method is its view-dependent nature (see Figure 6.7). A zoom operation reveals more details than an overview of a big image. Since the lighting configuration is calculated in screen-space, the scale of the details that are enhanced is automatically adjusted. As a side effect, the variation of the light direction on the surface of the object is somewhat bound in screen space: the more the user zooms in on a detail, the more the light directions bend to the local features. This is very important for very large images. When looking at the whole image, the light directions on the full-size original image could vary too fast, leading to significantly unpleasant visual effects.

(a) *Standard rendering*

(b) *Anisotropic sampling*

(c) *Isotropic sampling*

(d) *Anisotropic sampling w/o smoothing*

(e) *Isotropic sampling w/o smoothing*

Figure 6.3: *Dynamic Multi-Lighting Enhancement: sampling strategy comparison* ($\alpha = 0.70$).

6.2.2 Static Multi-Lighting Enhancement

The Dynamic Multi-Lighting Enhancement can be easily modified to produce an automatic high-contrast, well-illuminated image for stand-alone presentations, high-quality printing, or similar purposes. This new version of the algorithm is called Static Multi-Lighting Enhancement. The static version of the algorithm differs from the dynamic one in two aspects: the lighting field is calculated on the entire image, not only on I^*; all the hemisphere of possible light directions is sampled during the enhancement optimization step. In this way, hundreds of local lighting configuration are tested and the best one, according to the enhancement metric (Equation 6.2), is selected for the final rendering.

This modification addresses the problem that great lighting variations across the image can produce visual artifacts. The algorithm uses some constraints in the light setup generation to overcome this problem and uses all the possible lighting directions during the image enhancement and rendering phase. The constrained light setup is generated by means of a multi-resolution framework (see Figure 6.4).

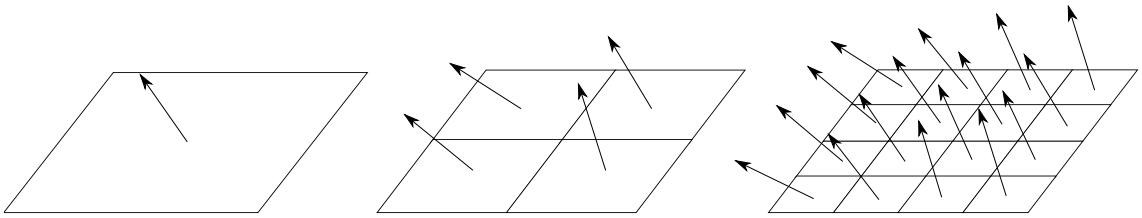


Figure 6.4: *Static multi-resolution lighting constraint. The light direction at the top level of subdivision influences the light directions at the successive levels of subdivision.*

At the first iteration the entire image is considered and all the possible light directions are evaluated (256 directions) according to the $\mathcal{E}$ metric (6.2) to estimate the optimal light direction (l'_0). At the second iteration the image is subdivided into four tiles and, for each tile, the optimal light is computed starting from l'_0. In general, after the first iteration the optimal light direction for each tile is chosen among a set of light vectors computed by perturbing the optimal direction of the parent tile with the sampling methods described for the dynamic enhancement (anisotropic or isotropic sampling). The process continues in the same way until it reaches a pre-defined level of subdivision. In order to reduce the computational cost of the algorithm the initial iterations works on a sub-sampled version of the image, using the appropriate mip-mapping level.

6.2.3 Results

The proposed Multi-Lighting Detail Enhancement algorithms can produce different visual effects thanks to the flexibility of the algorithm. The parameter α allows the

user to tune the enhancement metric to improve the contrast or the global brightness of the image. The Figure 6.5 shows an example where the parameter α increases the depth perception and the sharpness of the image and, at the same time, maintains a good global brightness (compare the faces of the human figures).

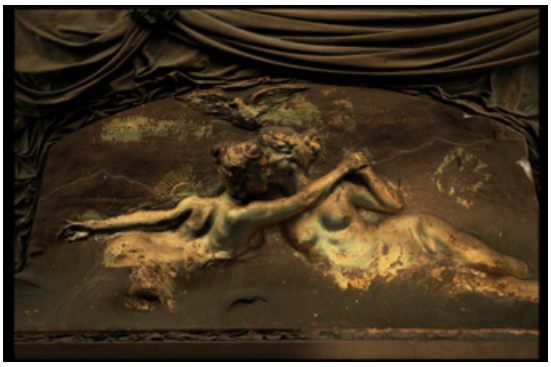
Figure 6.5: *(Left) Standard rendering. (Right) Dynamic Multi-Lighting Detail Enhancement with anisotropic sampling ($\alpha = 0.7$).*

Figure 6.6 shows the effects of the enhancement measure (Equation 6.2) varying the parameter α in the Static Multi-Lighting Enhancement.

Figure 6.8 shows a comparison of the tested sharpness operators M_1, M_2 and M_2 (Equation 6.5). In general, the Laplacian energy gives the best results because the second derivative is more accurate in the detection and extraction of edges and fine details. Hence, the final renderings present greater light variations. The L^1-norm and L^2-norm of the image gradient generate image with more uniform illumination. The L^2-norm can considerably amplify the details in some cases.

Another important parameter of Multi-Lighting Detail Enhancement is the light sampling strategy. The main difference among the two proposed methods is shown in Figure 6.3. With anisotropic sampling, the lighting tends to be grazing, because the light direction is constrained to rotate along the viewing axis; while with the isotropic sampling it becomes more perpendicular to the image surface. Furthermore, the two methods have opposite disadvantages: anisotropic sampling can generate excessively dark images in some parts; isotropic sampling can generate too bright images. These effects can be balanced out decreasing the α value for the former and increasing it for the later.

(a) *Standard rendering*



(b) *Static enhancement* $\alpha = 0.7$

(c) *Static enhancement* $\alpha = 0.4$

Figure 6.6: *Static Multi-Lighting Enhancement with anisotropic sampling. Image (b) has a high contrast but it is too dark due to the prevalence of contrast over brightness. Image (c) is well balanced (good contrast, good overall illumination).*

Performance

The proposed methods are tested on a commodity PC with CPU Intel Core 2 Quad Q9400 (2,67 GHz), 4.00 GB RAM and GPU NVIDIA GTX260 896MB, with three different PTM images: a sarcophagus in the monumental cemetery in Pisa (Figure 6.5) with resolution 2930×2224; a marble bas-relief in exposition at the Museum of the Opera del Duomo in Pisa (Figure 6.3) with resolution 3496×2280; and a high-relief in gilded wood representing a symbolic kiss between Corsica and Elba islands, from the Isola d'Elba Museum (Figure 6.6) with a resolution 2516×1646. All the computations are done in CPU taking advantage from a multi-core implementation. The rendering times are summarized in Table 6.1.

The Dynamic Multi-Lighting Enhancement works at interactive rates with an average rendering times of $140ms$. The rendering time is constants for all the test cases because the technique depends only on the view resolution. The viewport used in the tests was 1280×720 pixels with a 100% zoom factor. The methods was also tested with the same viewport but with a 51% zoom factor, the worst

Figure 6.7: *Dynamic Multi-Lighting Enhancement with anisotropic sampling. (Left) Standard rendering. (Right) Dynamic Detail Enhancement with* ($\alpha = 0.7$). *Changes in scale are automatically achieved thanks to the view-dependent nature of the algorithm.*

case for the selection of the mip-mapping level. In this case the rendering times is higher but it remains near interactive ($480ms$). On the contrary, the Static Multi-Lighting Enhancement depends on the image resolution and requires some seconds to generate the final static image that cannot be relighted. This technique is not interactive because it is designed for presentation purposes, in order to create an automatic high-contrast, well-illuminated image for stand-alone presentations or high-quality printing.

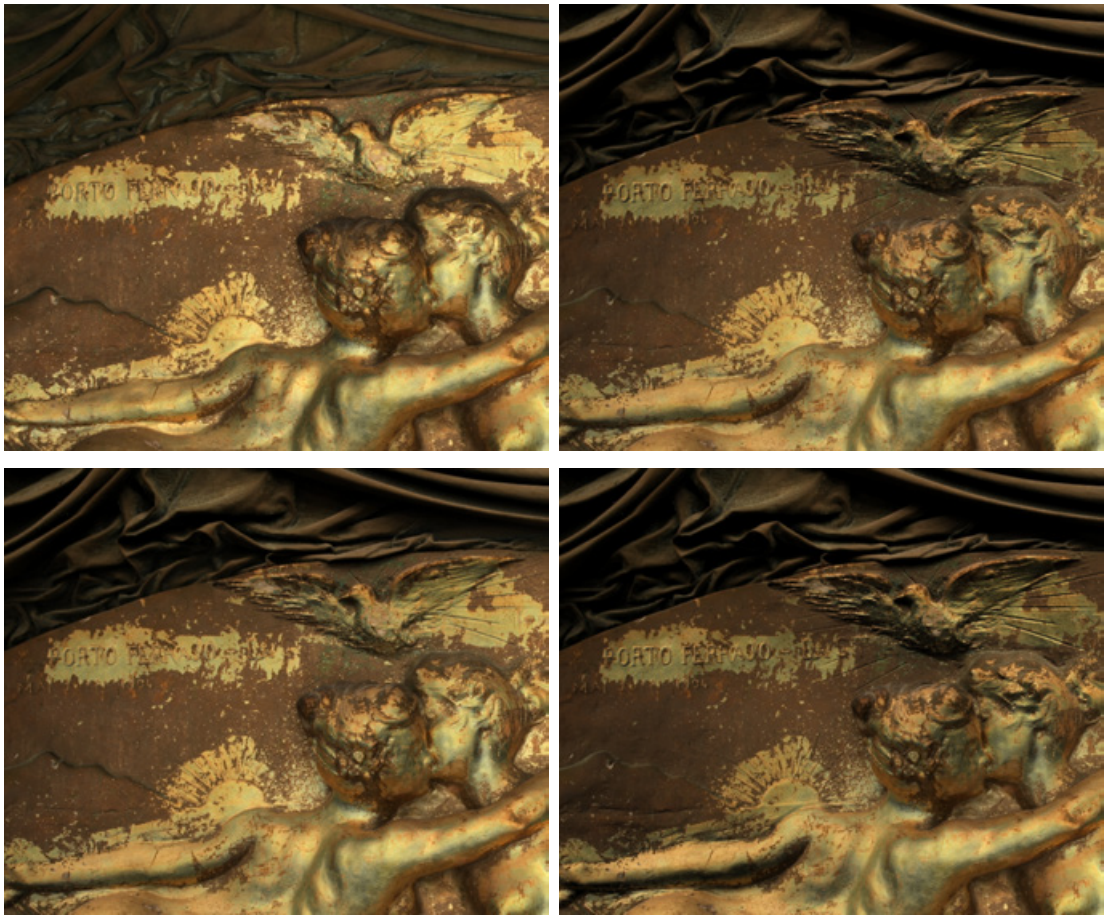


Figure 6.8: *Sharpness operator comparison in the Static Multi-Lighting Enhancement with $\alpha = 0.65$. (Top-left) Standard rendering. (Top-Right) Energy of Laplacian. (Bottom-Left) L^1-norm of the image gradient (Bottom-Right) L^2-norm of the image gradient.*

	Sarcophagus	Bas-relief	Corsica-Elba
Dynamic Enhancement	0.140s	0.140s	0.140s
Static Enhancement	7.03s	9.72s	4.38s

Table 6.1: *Multi-Lighting Detail Enhancement: performance.*

6.3 RTI Web Interactive Presentation

The recent advances of the web visualization instruments, such as the new JavaScript API WebGL that makes the integration and the use of the 3D content in the web browser more immediate and easy, can be exploited to increase the current capabilities to disseminate the RTI media and to support their remote visual inspection of both scholars and ordinary public.

This section presents a web viewer for RTI images that, taking advantage from the WebGL standard and from a multiresolution image format, allows the real-time interaction with the image using a set of basic functionalities (panning, zooming and changing the light direction). This viewer has been used and tested in the contest of a concrete and practical project: the development of an interactive kiosk for the presentation and virtual inspection of the coin collection of the National Museum of San Matteo in Pisa, using HTML5 technology and RTI images.

6.3.1 RTI Web Viewer

The developed viewer exposes a set of basic functionalities to interact with the RTI image, like the panning, the zooming and the changing of the light directions. It has been developed using WebGL and SpiderGL to allow the real-time rendering of high resolution RTI images in the modern web browsers with an efficient data loading. WebGL [102] is a library that extends the capability of JavaScript to allow the interactive generation of 3D content within any compatible web browser (Chrome, Firefox, Opera, Safari). The main advantage of WebGL, with respect to the previous solutions for the integration of 3D content in the web, is the absence of external plug-ins to be installed, because it is a build in feature of modern web browsers. SpiderGL [50] is a Computer Graphics JavaScript library that provides a set of data structures and algorithms to simplify the development of a WebGL application. It is composed by a number of modules to define and manipulate shapes, to import 3D models in several formats, to handle asynchronous data loading and to manage the user interaction. The use of WebGL and SpiderGL allows the customization of the Graphics Processor Unit pipeline directly from the web browser with specialized shaders, which permit a complete interaction with the RTI image changing even the light direction in real time.

In order to allow the user to interact immediately with the image, without the awaiting of the complete loading of the data, a new streamable multiresolution RTI format has been proposed. This encoding needs a hierarchical layout of the data, to prepare the image to be stored in a web server, an algorithm to visit such hierarchy and determine the nodes to use for producing the current viewport and the ability to load the nodes of the hierarchy asynchronously, proceeding with the rendering while missing data are fetched.

The procedure for the generation of the multiresolution format subdivides the RTI image in layers: nine layers for a HSH format (Figure 6.9); six layers for a RGB-PTM format; three layers for the LRGB format. Each layer stores three coefficients of the per-pixel reflectance functions: for the HSH and RGB-PTM the i-th layer stores i-th coefficient of the three RGB color channels; for LRGB-PTM the first two layers store the six coefficients of the biquadratic polynomial and the third one the RGB color. Then, for each layer, it creates a multiresolution tree and it cuts each level of the tree in tiles using a quad-tree structure (Figure 6.10). Finally it saves each tile in a different PNG image. This means that, to visualize a specific

pixel, all the PNG images that contain the coefficients of the pixel must be loaded. The PNG image format guarantees a lossless compression that reduces the amount of data to transfer from the server to the client without to lose any information needed for the photorealistic rendering. The main advantage of this multiresolution streamable format is the out-of-core loading of the tiles making immediately available at least some low resolution data and allowing to start quite instantaneously the user interaction with the image. In the specific at the beginning the user interacts with a low resolution version of the image, which is progressively refined as soon as the higher resolution data are loaded. The loading of the tiles at the different resolutions is guided by the zoom and pan operation of the user.

Figure 6.9: *Hemispherical Harmonics layer decomposition.*

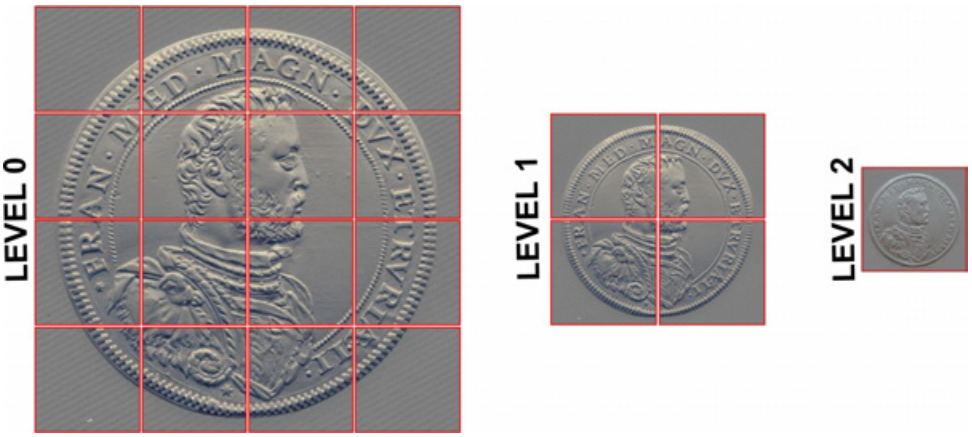


Figure 6.10: *Multi-resolution streamable quad-tree encoding.*

6.3.2 The San Matteo Coins Project

The San Matteo coins project started from the request of the curators of the National Museum of San Matteo in Pisa to present the ancient coins collection of the museum in an innovative way, in order to capture the interest of the visitors and to give them enhanced information. The attention of the museum curator towards the coin collection is due to the current way to expose and present it at the public. A coin is a very small artwork, which in a standard museum exposition is presented to the public from a distance (typically at least 50cm far from the observer eyes). This distance does not allow the visitors to note some small and interesting details on the legend or on bas-relief and the coin is usually visible only from one side. Furthermore, coins have a lot of hidden knowledge that is difficult to transfer to the visitors in an easy, effective and understandable manner. Then the main challenges of this project were:

- to allow the virtual manipulation of the coins to inspect them in detail;
- to bring some of the hidden knowledge of the coins to the ordinary public in an easy and understandable way.

The solution was the design and implementation of an interactive kiosk to allow the presentation and the virtual inspection of the coins collection. The kiosk must be easy and intuitive to use for the ordinary public of the museum and it must allow the real-time manipulation of the coin with a set of basic operations like zooming, panning, flipping the coin, changing of the light direction. Finally, the kiosk must tell the story of the coins using multimedia data, with a combination of text, images and videos. For this last purpose the coins are organized in several subsets, each one characterized by a feature that can be the historical collection to which the coins belong to or a common thematic subject (for example the coin of a specific geographic area or epoch). In addition, some hot spots are located on the surface of each coin with additional multimedia information useful to better understand the most important and interesting details depicted on the coin engraved decorations.

The management of the virtual inspection of the coins is one of the most important aspects in the design of the kiosk. Several scientific researches [161] [3] proved that the human brain is able to infer more cognitive data from the dynamic reflection and shading of an object. Then it is a fundamental requirement to give to the user the possibility to rotate the coin under, at least, a directional illumination. This means that the kiosk needs a virtual representation of the coin that can simulate the illumination effects in real-time and in an accurate way, in order to obtain photo-realistic renderings. The choice of this virtual representation is critical because the production of a photo-realistic rendering of a coin is extremely complex, due to the reflection effects of the different types of materials. For example in the museum collection there are both high reflective and specular gold coins and bronze coins, which are more opaque and have on their surface different kind of patinas and degradation processes that alter their appearance (Figure 6.11).

Figure 6.11: *Two example of coins selected for the kiosk: (Left) a Roman bronze coin; (Right) a modern gold coin.*

There are two possible options for the virtual representation: a complete 3D model or a RTI image. The creation of a 3D model requires acquiring both the 3D geometry and the surface appearance. This task could be quite complex. The acquisition of the geometry can present problems with the scanning of the coins border, that can be very thin, and the right alignment of the two sides of the coin. The acquisition of the reflectance must be done in a following step with special setups to sample in an accurate way its dimensionality. Even if several solutions have been proposed for the acquisition of the surface appearance of real object, all ones have some drawbacks. They require a very intensive data acquisition, which must sample in accurate way both the light and the view directions, and complex reflectance models that, due to the assumptions on the reflection effects that we want to capture, are not able to reproduce all type of materials. Finally, the processing of the acquired data for the creation of the final 3D model is time consuming and the manipulation of a 3D model is still more complex to understand and control for the user.

On the other hand, RTI techniques produce a 2D representation of each coin that encodes both the surface normal and the appearance in a single image. This image can be dynamically relighted by the user reproducing the illumination-dependent effects of the surface with a higher quality and a higher resolution that is not usually provided by 3D scanned model. The acquisition and processing step are cheaper than 3D scanning and the final representation simplify the interaction because the user is more accustomed to interact with an image than with a 3D model. For these reasons, RTI images were chosen for the realization of the San Matteo kiosk.

Acquisition and Processing of the RTI data

The first step of the project was the acquisition and the generation of the RTI images. For the acquisition, the museum curator selected a subset of 41 coins from the museums collection, following value and storytelling criteria. The coins cover different epochs, from the Roman Empire to the Grand Duchy of Tuscany (XVI - XIX centuries).

The digitization was done with the minidome designed by the University of Leuven [206]. The dome is composed by 4 shells that can easily assemble and disassemble to simplify the transport. It has 260 white LEDs and an overhead CCD camera. The device is computer controlled to allow a completely automatic acquisition. For each coin, the acquisition takes about 10 minutes, required to shot and store 520 photos (260 photos for each side). All the coins were digitized in a single working day.

The processing of the acquired raw data to produce the final RTI image involved 3 steps:

- the transformation of the raw images taken with the minidome from the Bayer Pattern to a RGB format;
- the generation of an RTI image for each coin side using the tools provided by Cultural Heritage Imaging corporation [36];
- the generation of the multiresolution streamable RTI format.

The Hemispherical Harmonics (HSH) format was chosen for the RTI images because it guarantees a better reproduction of the specular reflection with results that are more photorealistic with respect to a Polynomial Texture Map, as showed in [135].

The generation of the RTI images took about 24 hours of automatic processing, without user intervention.

The Interactive System

The system is composed by two integrated sections. The first section allows the presentation to the user of the different subsets of coins. The second section permits the interactive RTI visualization. The main features of the kiosk are:

- the organization of the coins in categories and the presentation of these categories with multimedia data;
- the virtual inspection of the coin by RTI manipulation. A general presentation and some hot-spots, that are located on selected areas on the surface of the coin to tell the most important and significant details, are associated to each coin;

- the possibility to run the kiosk on a web site or on a touch screen system (an interactive installation inside the museum) thanks to the technologies used for the development: HTML and JavaScript for the general structure of the kiosk and for the presentation of multimedia data; WebGL and SpiderGL for the RTI visualization.

The Figures 6.12, 6.13, 6.14 and 6.15 show some screenshots of the kiosk. Entering from the cover page (Figure 6.12 top-left) the user has a presentation page of the project and, on the left, a menu to navigate among the different pages of the first section and to access directly the RTI viewer (Figure 6.12 top-right). The items in the menu allow the access to the different subsets of coins. The subsets are subdivided into two categories: the historical collections (Figure 6.12 bottom-left) and the thematic subject (Figure 6.12 bottom-right). By selecting a subset, the user can read additional information about it (Figure 6.13). At the left of the page there is a scrollable bar with the thumbnail and the name of the coins in the subset. By clicking on a coin in this bar, the user can open the visualization of the relative RTI image. In the page of RTI image (Figure 6.14), there is the RTI viewer in the middle, a general description of the coin on the left and the scrollable bar on the right. Using this bar the user can switch very quickly on other coins of the current subset. The RTI viewer has a title bar on the top with the name of the coin and a tool bar on the bottom that allows the manipulation of the images. The user can change the light direction, pan the image, zoom in and zoom out, flip the coin to see the other side, enable the visualization of the hot spots. By clicking on a hot spot the user can display the relative multimedia content on the left side of the page (Figure 6.15). The visualization of the hot spots content is preceded by an automatic zoom animation to better highlight the detail associated to it. The arrows in the bottom of the dialog allow the scrolling of the different images associated with the hot-spot.

An interactive kiosk will be installed at the end of 2013 at the National Museum of San Matteo in Pisa. The installation setup will be composed by a 24-inch multitouch screen, used for the user input and visualization, paired by a bigger screen, set on the side, that shows the same content of the touch screen, allowing a clean vision for the other visitors that do not interact directly with the kiosk. To understand the real effectiveness of the current organization of the kiosk, the system is provided with a data logging framework that allows recording the interaction of the users with the system, that can be used in following analysis.

Figure 6.12: . *(Top-left) Cover page. (Top-right) Presentation page. (Bottom-left) The historical collections. (Bottom-right) The thematic subject.*

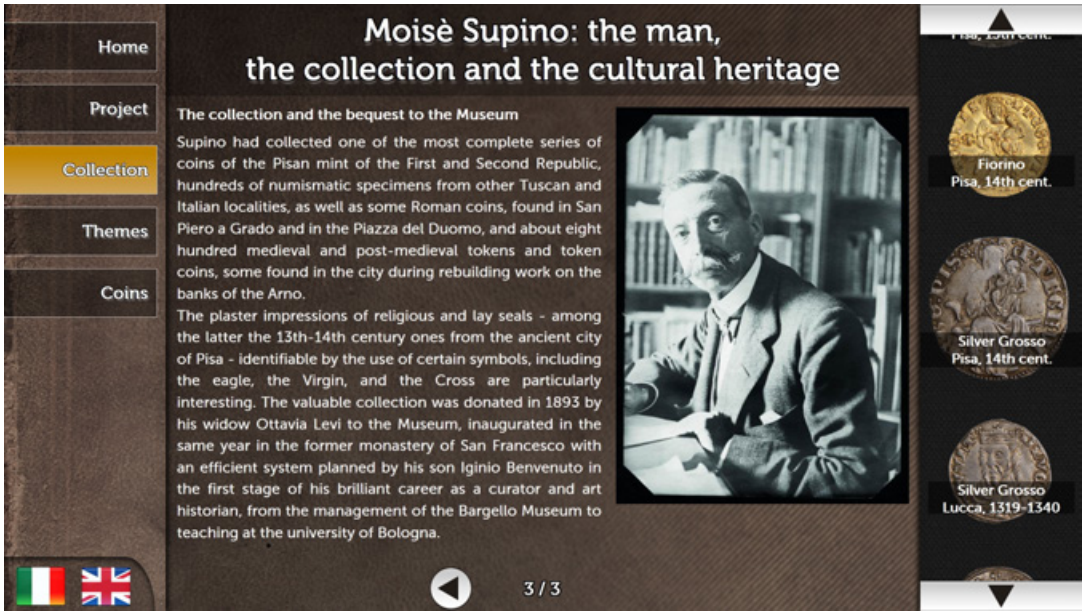


Figure 6.13: *Category content.*

Figure 6.14: *RTI viewer: (Top-left) starting page; (Top-right) activation of the hot-spots; (Bottom-left) coin detail; (Bottom-right) coin detail presented under a different light direction.*

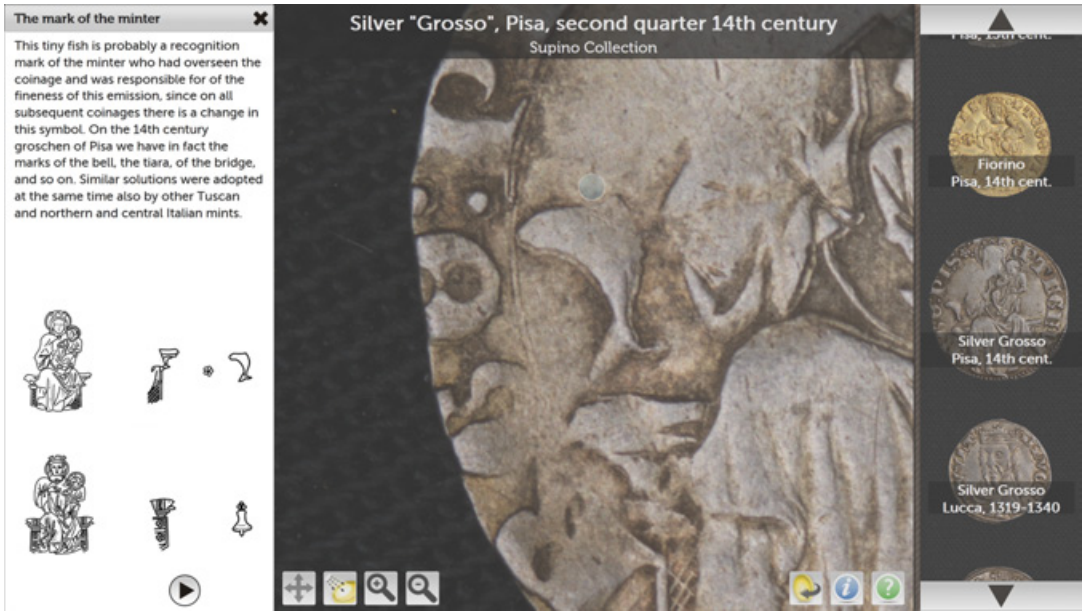


Figure 6.15: *Visualization of the content activated by a hot-spot.*

Chapter 7

Conclusion

The acquisition and the visualization of the surface appearance is a challenging and important task for the virtual representation of Cultural Heritage artworks. This thesis has proposed some innovative approaches: an algorithm for the accurate and efficient alignment of a video sequence of a real object over its 3D model; two new algorithms for the estimation of the Surface Light Field and Spatially Varying BRDF from video sequences acquired in general and fixed lighting condition; a shading detail enhancement for RTI; a web interactive system that allows the presentation and the virtual inspection of artwork collection using HTML5 and RTI images, tested in a concrete digitization project. This chapter summarizes these works and proposes some possible future extensions.

7.1 Appearance Estimation from Video Sequences

The acquisition and estimation of the surface appearance is a challenging activity in the Cultural Heritage field for several reasons: objects of very different scale, with different type of materials, even on the same objects, to acquire in few controllable lighting environments. Unfortunately, the available solutions show several limitations: lack of flexibility because developed to solve very specific problems; acquisition in lab with controlled lighting conditions, difficult to reproduce in an on-the-field acquisition; high time to acquire, calibrate and elaborate huge sets of images; high level of expertise to evaluate the completeness of the photographic acquisition.

This thesis proposed two new methods to compute two different appearance approximations, the Surface Light Field and the Spatially Varying BRDF, starting from video sequences made easy and acquired in the natural environment in which the object is placed. Both the methods are based on the same framework, composed by two stages:

- the registration of the video sequences over the mesh by calibration of the camera, to compute for each video frame the camera parameters that describe

the projection process of a 3D point into the image plane;

- the reconstruction of the surface appearance based on the separate estimation of the reflectance components, starting from a first estimation of the acquisition lighting environments.

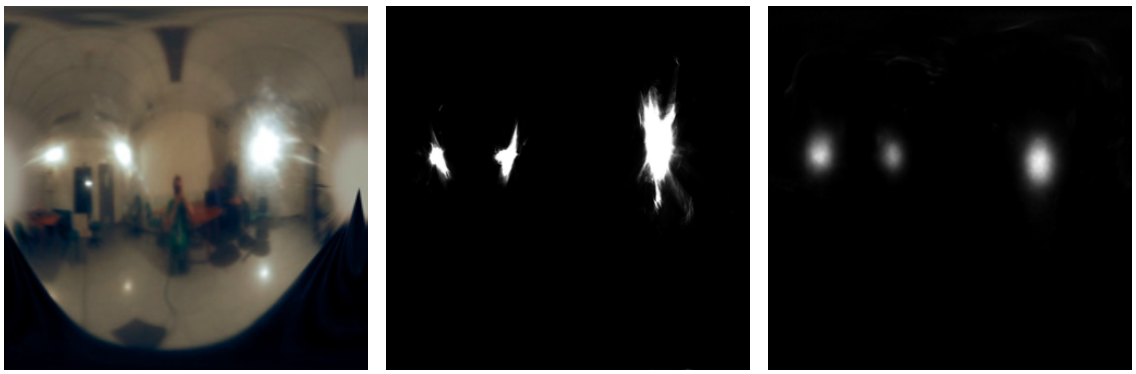
Chapter 3 presented a new algorithm for the registration of a video sequence of a real object over its dense digital representation, taking advantage of the high frame-to-frame coherence. It puts together the strong-points of two different alignment approaches: feature-based by KLT video tracking; statistical-based by maximizing the MI between the gradient map of the frames and the gradient map of the rendering of the 3D model with two illumination related properties, the surface normal and the ambient occlusion. The registration by MI is able to correct the drift problem introduced by the KLT tracker in long and complex sequences, while KLT tracking speeds up the registration and controls the convergence of MI towards good camera parameters. The accuracy of the registration was tested on a synthetic sequence and on four real videos of object of different size. Results are extremely positive, with a very low projection error. This makes the algorithm useful in all the applications that need the bi-directional data transfer between the 3D model and the video, like color transfer, estimation of reflectance properties and recording of appearance-varying scenes.

Chapter 4 presented a new method for the estimation of the Surface Light Field starting from a 3D scanning model and some video sequences made easy, acquired moving the camera around the object. The input videos do not guarantee a uniform sampling density of the view direction. In order to avoid interpolation artifacts due to the very irregular and not uniform video acquisition, characterized by a dense coverage sampling only along the camera path, the method separates the SLF in two components: the diffuse color and the residual from the diffuse color. The first component is estimated using statistical operations that take advantage of the data redundancy of the video sequences. The main idea is to discard all the color samples that have a higher probability to exhibit a not diffuse behavior. It roughly estimates the direction of the main light sources by accumulation of the saturated samples along the specular mirror directions, and then it deletes all the samples that have an angle between the view direction and the specular mirror direction of the light vector above a fixed threshold. The second component models the color residual, which is the difference from the diffuse color, as linear combination of a basis of spherical functions. The results on the test cases do not present artifacts due to the interpolation and fitting of the spherical functions and the similarity measures proof a high fidelity degree between the renderings with the estimated SLF and the original video frames used by the algorithm. Finally, a small change in the rendering formula was introduced to enhance the residual component, in order to overcome the limitation of a band-limited fitting and, at the same time, to preserve the real-time visualization of the model.

Chapter 5 presented a statistical method for the acquisition of the Spatially Varying BRDF of complex object starting from video sequences taken under fixed and uncontrolled lighting conditions. Given the video frames and a 3D model of the object, the method is able to estimate the environment map of the scene, using the object itself as a probe and capturing enough lighting details for the Spatially Varying BRDF estimation, a good approximation of the diffuse color, without view depended reflection effects, and the specularity parameters of the basis materials, segmented on the object in an assisted way from the user. Given the limited input data and the very easy acquisition process, the results show that, even in the case of complex and multi-material objects, the reflectance properties are estimated with an accuracy that produces very realistic renderings. Although the method presents some limitations (due to the type of input data and the specularity model applied to describe the materials), the trade-off between the easy of acquisition and the obtained results makes it extremely useful for practical applications. This is especially true when an on-the-field acquisition has to be performed, and the interaction with the object and the surrounding environment is limited.

The two methods proposed in the Chapters 4 and 5 are based on a common pipeline: reconstruction of the acquisition lighting environment; estimation of the diffuse color; modeling of the other reflectance effects. Especially the first two stages present some similarities. Both the methods try to estimate the acquisition lighting configuration through the construction of an approximated environment map by projection along the specular mirror direction, which is following used to improve the computation of the diffuse color. In the Chapter 4 the method uses a simple luminance thresholding to select all the color samples that are acquired in the saturation areas of the camera CCD. These samples have a higher probability to show a not diffuse behavior. This environment maps is following approximated via clustering with a set of directional lights that are used for the estimation of the diffuse colors. This estimation uses a fixed angular threshold to discard all the samples too near at the mirror direction of the light vectors. The angular threshold is not chosen to obtain a right classification of the color samples but to decrease the statistical influence of the samples with a higher probability to exhibit a not diffuse behavior in the computation of the diffuse color. In this process, it takes advantage from the availability of a high number of color samples projected for each surface point. On the other hand the Chapter 5 proposed a more automatic method based on the computation of a per-point adaptive thresholds, based on the projected color samples. Then the environment maps is used to computed a new quality weight for each samples that make more robust the estimation of the diffuse color. The obtained results are quite similar with some differences. Figure 7.1 shows a comparison of the environment maps created by the two methods proposed in the Sections 4.2.2 and 5.4, using the same dataset (the DWARF video). Both ones are able to approximate the position of the main light sources, with the advantage that the method in the Section 5.4 produces less noisy results. Figure 7.2 shows a comparison of the diffuse color estimated by the method described in the Sections

4.2.3 and 5.5. The results are very similar with very small differences, due to small highlights that are not correctly removed.



(a) (b) (c)

Figure 7.1: *Environment maps comparison: (a) real environment map; (b) environment map estimated with the method described in Section 4.2.2; (c) environment map estimated with the method described in Section 5.4.*

(a) (b) (c) (d)

Figure 7.2: *Diffuse color estimation comparison: (a) and (c) estimated with the method described in the Section 4.2.3; (b) and (d) estimated with the method described in the Section 5.5.*

7.2 Reflectance Transformation Imaging

Although RTI techniques allow a partial reconstruction of the surface appearance, limiting the interaction to the only light direction, this technology is widely used in the Cultural Heritage context for the acquisition of near flat objects for several reasons: inexpensive and widely available hardware; simple acquisition; scale well with the size of the artwork; high sampling density and precision; short processing time; photo-realistic rendering of challenging materials.

As parallel work, this thesis presented two solutions to address two important open issues in this context: how to provide better and more flexible virtual inspection capabilities with a set of operators that improve the perception of details, features and overall shape of the artwork; how to increase the possibility to disseminate this data and to support remote visual inspection of both scholar and ordinary public.

Chapter 6 presented a new shading enhancement method focused on the task of locally optimizing the light direction to improve the sharpness and brightness of the resulting final image. The main idea is to find for each region of the image the light direction that maximizes an enhancement metric, which is a linear combination of the local sharpness and brightness. Two versions of the algorithm have been proposed. The dynamic version perturbs locally the main light direction chosen by the user in a view dependent way, allowing a real-time interaction. The static version explores all the hemisphere of possible light directions to produce a single well-illuminated static image that can be used for presentation purposes. It uses multi-resolution constraints in the generation of the virtual light setup. The second part of the chapter presented a web interactive viewer for RTI images developed using HTML5 and WebGL. The viewer has been used as central and innovative feature in the development of a museum kiosk for the presentation of artworks collections. The kiosk, deployable even remotely in a web site, is composed by two sections. The first section allows the introduction and the presentation of the artworks using multimedia data and several thematic paths. The second section permits the interactive RTI visualization with a set of basic operation, like zoom, pan, the change of the light direction and the visualization of the hot spots. The system has been used in the contest of a concrete and practical project, the digitization of the coin collection of the National Museum of San Matteo in Pisa.

7.3 Future Work

Even if the final results are effective and realistic, there are still open issues that can be investigated to further improve the proposed techniques.

The video-to-geometry registration algorithm has shown a limitation in the case of major occlusions, like the ones that show-up in the sequence of the Nettuno statue (due to the occlusions with the other statues that are part of this complex monument). In this case, an improvement in the registration can be obtained or with

an automatic procedure that deletes the 2D features on the occluders taking into account several info, like the camera motion or the error returned by the algorithm for each 2D-3D correspondence, or implementing a multi-step registration algorithm with several application of forward and backward registration. A further improvement could be the possibility to make the entire algorithm completely automatic, removing the need of an initial manual alignment of the first frame.

The proposed algorithms for the estimation of the appearance have several common future working directions. The first one is to study the robustness of the algorithms changing the temporal sampling and the image resolution, for example comparing the results produced on video data with the ones produced on photo sequences acquired with a lower frame-rate but with a higher resolution, moving the camera while continuously shooting around the object. The second one is the improvement of the rendering results of objects with a significant meso-structure reconstructing an optimized additional bump map. Then, it would be interesting to test the methods with 3D models created directly from the input video sequences, using a 3D multi-view reconstruction. In this case, the quality of the output model from multi-view stereo is a critical point in the estimation of the reflectance due to the accuracy of the computed surface normals.

Furthermore, in order to improve the quality of the final Surface Light Field rendering and to get over the band-limited reconstruction of the specularity, it should be interesting to estimate in an automatic way the enhancement parameter I_s in Equation 4.11. While the current modeling of the specular component with spherical functions is done independently for each surface point exploring only the temporal coherence of the video (using the color samples projected on the point by each frame), a future work could be exploring the spatial consistency of some selected frame to set an optimization procedure for the computation of the best I_s value.

For the approximation of the Spatially Varying BRDF, two other improvements can be further investigated: to use a more complex BRDF model to reproduce in a more accurate way the reflectance properties of the materials; to extend the algorithm to use HDR videos acquired with a professional HDR camera, in order to obtain more accurate and reliable data about the specular reflection.

In the RTI context the future working directions are the extension of Multi-Lighting Detail Enhancement to 3D models, taking advantage from the mathematical model used to model the surface appearance, and the study of the real effectiveness of the interactive web presentation system, analyzing the data stores by the logging framework. From this analysis it should be possible to evaluate different things, such as the easy of the GUI, the relation of the user with the new RTI technology (time spent in the interactive manipulation session), simple statistics on the most viewed categories and coins, and so on.

7.4 List of Publications

The research contribution presented in this thesis was the subject of the following publications (in chronological order):

- Gianpaolo Palma, Massimiliano Corsini, Paolo Cignoni, Roberto Scopigno, Mark Mudge
  Dynamic Shading Enhancement for Reflectance Transformation Imaging
  ACM Journal on Computing and Cultural Heritage, Volume 3, Issue 2, pp. 6:1-6:20, 2010

- Mark Mudge, Carla Schroer, Graeme Earl, Kirk Martinez, Hembo Pagi, Corey Toler-Franklin, Szymon Rusinkiewicz, Gianpaolo Palma, Melvin Wachowiak, Michael Ashley, Neffra Matthews, Tommy Noble, Matteo Dellepiane
  Principles and Practices of Robust, Photography-based Digital Imaging Techniques for Museums
  VAST 2010, The 11th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage - Short and Project Papers, Eurographics Association, pp. 111-137, 2010

- Gianpalo Palma, Marco Callieri, Matteo Dellepiane, Massimiliano Corsini, Roberto Scopigno
  Geometry-aware Video Registration
  VMV 2010: Vision, Modeling and Visualization Workshop, Eurographics Association, pp. 107-114, 2010

- Gianpaolo Palma, Massimiliano Corsini, Matteo Dellepiane, Roberto Scopigno
  Improving 2D-3D Registration by Mutual Information using Gradient Maps
  Eurographics Italian Chapter Conference 2010, Eurographics Association, pp. 89-94, 2010.

- Gianpalo Palma, Eliana Siotto, Marc Proesmans, Monica Baldassarri, Clara Baracchini, Sabrina Batino, Roberto Scopigno
  Telling the Story of Ancient Coins by Means of Interactive RTI Images Visualization
  CAA 2012: Computer Applications and Quantitative Methods in Archaeology, 2012, (To appear)

- Gianpaolo Palma, Marco Callieri, Matteo Dellepiane, Roberto Scopigno
  A Statistical Method for SVBRDF Approximation from Video Sequences in General Lighting Conditions
  Computer Graphics Forum (Proceeding of the 23th Eurographics Symposium on Rendering), Volume 31, Issue 4, pp. 1491-1500, 2012

- Jaime Kaminski, Karina Rodriguez Echavarria, David Arnold, Gianpaolo Palma, Roberto Scopigno, Marc Proesmans, James Stevenson
  Insourcing, Outsourcing and Crowdsourcing 3D Collection Formation: Perspectives for Cultural Heritage Sites
  VAST 2012, The 13th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage, Eurographics Association, pp. 81-88, 2012

- Gianpaolo Palma, Nicola Desogus, Paolo Cignoni, Roberto Scopigno
  Surface Light Field from Video Made Easy
  (Submitted to Digital Heritage 2013).

Appendix A

SVBRDF Statistical Estimation: Math Background

A.1 Median Upper Bound

Let μ, m and the σ be respectively the mean, the median, and the absolute deviation of a random variable X. Then

$$|\mu - m| < \sigma. \tag{A.1}$$

Proof. From the Jensen's inequality we know that if X is a random variable and φ is a convex function, then $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$. Then we have

$$\begin{aligned} |\mu - m| &= |\mathbb{E}[X] - m| && \text{(A.2)} \\ &= |\mathbb{E}[X - m]| && \text{(A.3)} \\ &\leq \mathbb{E}[|X - m|] && \text{(A.4)} \\ &\leq \mathbb{E}[|X - \mu|] = \sigma && \text{(A.5)} \end{aligned}$$

where the first inequality comes from the Jensen's inequality applied to the absolute value function, which is convex. The second inequality is true because the median minimizes the absolute deviation. □

A.2 Specular Parameters Computation

Given the unknown ρ_s and α, the following system of equations:

$$\begin{cases} \rho_s \; A^\alpha &= L_A - L_d \\ \rho_s \; B^\alpha &= L_B - L_d \end{cases} \tag{A.6}$$

can be solved by applying the logarithm to both the equations:

$$\begin{cases} \alpha = \dfrac{(\ln(L_A - L_d) - \ln(\rho_s))}{\ln(A)} \\ \alpha = \dfrac{(\ln(L_B - L_d) - \ln(\rho_s))}{\ln(B)} \end{cases} \tag{A.7}$$

and by setting the equality between the two equations:

$$\frac{(\ln(L_A - L_d) - \ln(\rho_s))}{\ln(A)} = \frac{(\ln(L_B - L_d) - \ln(\rho_s))}{\ln(B)} \tag{A.8}$$

From equation A.8 we compute the value:

$$\rho_s = e^C \tag{A.9}$$

where:

$$C = \frac{\ln(L_B - L_d)\ln(A) - \ln(L_A - L_d)\ln(B)}{\ln(A) - \ln(B)} \tag{A.10}$$

With the value of ρ_s we recover the value of α by solving one of the equation in the system A.7.

Bibliography

[1] Y. Abdel-Aziz and H. Karara. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. In *ASP Symposium on Close Range Photogrammetry*, pages 1–18, 1971. 12, 13

[2] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20, 1991. 38

[3] E H Adelson and A P Pentland. The perception of shading and reflectance. *Perception*, 1:409–423, 1996. 124

[4] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *ACM Transactions on Graphics (TOG)*, 23(3):294–302, 2004. 44

[5] D. Akers, F. Losasso, J. Klingner, M. Agrawala, J. Rick, and P. Hanrahan. Conveying Shape and Features with Image-Based Relighting. *Proceedings of the 14th IEEE Visualization 2003 (VIS'03)*, pages 349–354, 2003. 44

[6] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. FREAK: Fast retina keypoint. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517. IEEE, 2012. 21

[7] N. Alldrin, T. E. Zickler, and D. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*, pages 1–8, 2008. 37

[8] Michael Ashikhmin and Peter Shirley. An anisotropic Phong BRDF model. *Journal of Graphics Tools: JGT*, 5(2):25–32, 2000. 30

[9] Shaun Bangay and Judith D. Radloff. Kaleidoscope configurations for reflectance measurement. In *AFRIGRAPH '04: Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, pages 161–170. ACM, 2004. 31

[10] João Garcia Barbosa, João Luís Sobral, and Alberto José Proença. Imaging techniques to simplify the ptm generation of a bas-relief. In *The 8th International Symposium on Virtual Reality Archaeology and Cultural Heritage VAST 2007*, pages 28–31, 2007. 43

[11] Adam Baumberg. Blending images for texturing 3D models. In *Proceedings of the British Machine Vision Conference 2002, BMVC 2002*, pages 404–413. British Machine Vision Association, 2002. 46

[12] Herbert Bay, Tinne Tuytelaars, and Luc J. Van Gool. SURF: Speeded up robust features. In *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Proceedings*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer, 2006. 20, 21

[13] Etienne Beauchesne and Sébastien Roy. Automatic relighting of overlapping textures of a 3D model. In *Computer Vision and Pattern Recognition*, pages 166–176. IEEE Computer Society, 2003. 45

[14] Fausto Bernardini, Ioana M. Martin, and Holly Rushmeier. High-quality texture reconstruction from multiple scans. *IEEE Transactions on Visualization and Computer Graphics*, 7(4):318–332, October 2001. xi, 46

[15] C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 1, pages 55–58. Springer, 2006. 55

[16] James F. Blinn. Models of light reflection for computer synthesized pictures. In *Computer Graphics (SIGGRAPH '77 Proceedings)*, volume 11, pages 192–198, July 1977. 30

[17] Samuel Boivin and André Gagalowicz. Image-based rendering of diffuse, specular and glossy surfaces from a single image. In *SIGGRAPH 2001 Conference Proceedings, August 12–17, 2001, Los Angeles, CA*, pages 107–116. ACM Press, 2001. 32

[18] Louis Borgeat, Guy Godin, François Blais, Philippe Massicotte, and Christian Lahanier. GoLD: interactive display of huge colored and textured models. *ACM Transactions on Graphics*, 24(3):869–877, July 2005. 45

[19] Alexander Bornik, Konrad F. Karner, Joachim Bauer, Franz Leberl, and Heinz Mayer. High-quality texture reconstruction from multiple views. *Journal of Visualization and Computer Animation*, 12(5):263–276, 2001. 45

[20] M. Brown and D. G. Lowe. Unsupervised 3D object recognition and reconstruction in unordered datasets. In *3DIM '05*, pages 56–63, 2005. 24

[21] Lionel Brunie, Stéphane Lavallée, and Richard Szeliski. Using force fields derived from 3D distance maps for inferring the attitude of a 3D rigid object. In *Computer Vision - ECCV'92, Second European Conference on Computer Vision, Proceedings*, volume 588 of *Lecture Notes in Computer Science*, pages 670–675. Springer, 1992. 16

[22] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *SIGGRAPH 2001 Conference Proceedings*, pages 425–432, 2001. 41

[23] Marco Callieri, Paolo Cignoni, Massimiliano Corsini, and Roberto Scopigno. Masked photo blending: mapping dense photographic dataset on high-resolution 3d models. *Computer & Graphics*, 32(4):464–473, Aug 2008. xi, 46, 47, 77, 93

[24] Marco Callieri, Paolo Cignoni, and Roberto Scopigno. Reconstructing textured meshes from multiple range RGB maps. In *Proceedings of the Vision, Modeling, and Visualization Conference 2002 (VMV 2002), November 20-22, 2002*, pages 419–426. Aka GmbH, 2002. 45

[25] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary robust independent elementary features. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pages 778–792. Springer, 2010. 21

[26] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH-00)*, pages 307–318, July 23–28 2000. 41

[27] Wei-Chao Chen, Jean-Yves Bouguet, Michael H. Chu, and Radek Grzeszczuk. Light field mapping: efficient representation and hardware rendering of surface light fields. *ACM Trans. Graph.*, 21(3):447–456, July 2002. 41

[28] O. Chum and J. Matas. Matching with PROSAC: Progressive sample consensus. In *CVPR*, pages I: 220–226, 2005. 14

[29] O. Chum, J. Matas, and J. V. Kittler. Locally optimized RANSAC. In *DAGM*, pages 236–243, 2003. 14

[30] Ioan Cleju and Dietmar Saupe. Stochastic optimization of multiple texture registration using mutual information. In *Pattern Recognition, 29th DAGM Symposium, Heidelberg, Proceedings*, volume 4713 of *Lecture Notes in Computer Science*, pages 517–526. Springer, 2007. 18

[31] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. In *Computer Graphics (SIGGRAPH '81 Proceedings)*, volume 15, pages 307–316, August 1981. 30

[32] Kurt Cornelis, Frank Verbiest, and Luc J. Van Gool. Drift detection and removal for sequential structure from motion algorithms. *IEEE Transactions on PAMI*, 26(10):1249–1259, 2004. 22

[33] Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, Riccardo Gherardi, Andrea Fusiello, and Roberto Scopigno. Fully automatic registration of image sets on approximate geometry. *International Journal of Computer Vision*, 2012. xi, 26, 27, 50

[34] Massimiliano Corsini, Matteo Dellepiane, Federico Ponchio, and Roberto Scopigno. Image-to-geometry registration: a mutual information method exploiting illumination-related geometric properties. *Computer Graphics Forum*, 28(7):1755–1764, 2009. xi, 18, 49, 51, 53, 57, 58, 64, 65, 67, 68, 69

[35] Yan Cui, Nils Hasler, Thorsten Thormählen, and Hans-Peter Seidel. Scale invariant feature transform with irregular orientation histogram binning. In *Image Analysis and Recognition, 6th International Conference, ICIAR 2009. Proceedings*, volume 5627, pages 258–267. Springer, 2009. 21

[36] Cultural Heritage Imaging. RTI Software. `http://culturalheritageimaging.org/What_We_Offer/Downloads/index.html`, 2012. [Accessed on May 2013]. 126

[37] Amaury Dame and Éric Marchand. Optimal detection and tracking of feature points using mutual information. In *ICIP*, pages 3601–3604, 2009. 22

[38] Kristin J. Dana, Shree K. Nayar, Bram van Ginneken, and Jan J. Koenderink. Reflectance and texture of real-world surfaces authors. In *CVPR*, pages 151–157. IEEE Computer Society, 1997. 33

[39] Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. Reflectance and texture of real-world surfaces. In *ACM Transactions on Graphics*, volume 18 (1), pages 1–34, 1999. 32

[40] P. David, D. F. DeMenthon, R. Duraiswami, and H. Samet. SoftPOSIT: Simultaneous pose and correspondence determination. *International Journal of Computer Vision*, 59(3):259–284, September 2004. 13

[41] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured Light Fields. *Computer Graphics Forum*, 31(2):305–314, 2012. 42

[42] L. S. Davis and D. F. DeMenthon. Model-based object pose in 25 lines of code. In *Image Understanding Workshop*, pages 753–761, 1992. 13

[43] Paul Debevec. A median cut algorithm for light probe sampling. In *ACM SIGGRAPH 2005 Posters*, 2005. 97, 102

[44] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 33

[45] Paul E. Debevec. Digitizing the parthenon: Estimating surface reflectance properties of a complex scene under captured natural illumination. In *Proceedings of the Vision, Modeling, and Visualization Conference 2004 (VMV 2004)*, page 99. Aka GmbH, 2004. 37

[46] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *SIGGRAPH 96 Conference Proceedings*, pages 11–20, 1996. 41, 46

[47] Matteo Dellepiane, Marco Callieri, Massimiliano Corsini, Paolo Cignoni, and Roberto Scopigno. Improved color acquisition and mapping on 3d models via flash-based photography. *ACM Journ. on Computers and Cultural heritage*, 2(4):1–20, Feb. 2010. 48

[48] Matteo Dellepiane, Massimiliano Corsini, Marco Callieri, and Roberto Scopigno. High quality ptm acquisition: Reflection transformation imaging for large objects. In *The 7th International Symposium on Virtual Reality Archaeology and Cultural Heritage VAST 2006*, pages 179–186. Eurographics Association, 2006. 43

[49] Matteo Dellepiane, Ricardo Marroquim, Marco Callieri, Paolo Cignoni, and Roberto Scopigno. Flow-based local optimization for image-to-geometry projection. *IEEE Trans. Vis. Comput. Graph*, 18(3):463–474, 2012. xi, 47

[50] Marco Di Benedetto, Federico Ponchio, Fabio Ganovelli, and Roberto Scopigno. SpiderGL: a JavaScript 3D graphics library for next-generation WWW. *Proceedings of the 15th International Conference on Web 3D Technology*, 1(212):165–174, 2010. 122

[51] Yue Dong, Jiaping Wang, Xin Tong, John Snyder, Yanxiang Lan, Moshe Ben-Ezra, and Baining Guo. Manifold bootstrapping for SVBRDF capture. *ACM Trans. Graph*, 29(4), 2010. 38

[52] Fadi Dornaika and Chistophe Garcia. Robust camera calibration using 2d to 3d feature correspondences. In *Proceedings of International Symposium SPIE*

– *Optical Science Enginnering and Instrumentation, Videometrics V, Volume 3171*, pages 123–133, 1997. 12

[53] Julie Dorsey, Holly Rushmeier, and François X. Sillion. *Digital Modeling of Material Appearance*. Morgan Kaufmann, 2007. 27, 29

[54] Martin Eisemann, Bert De Decker, Marcus Magnor, Philippe Bekaert, Edilson de Aguiar, Naveed Ahmed, Christian Theobalt, and Anita Sellent. Floating textures. *Computer Graphics Forum (Proc. of Eurographics)*, 27(2):409–418, April 2008. 47

[55] Raanan Fattal, Maneesh Agrawala, and Szymon Rusinkiewicz. Multiscale shape and detail enhancement from multi-light image collections. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007)*, 26(3):to appear, 2007. 44

[56] O. D. Faugeras and G. Toscani. The calibration problem for stereo. In *Proceedings, CVPR '86 (IEEE Computer Society Conference on Computer Vision and Pattern Recognition)*, pages 15–20. IEEE, 1986. 12

[57] Olivier D. Faugeras, Quang-Tuan Luong, and Stephen J. Maybank. Camera self-calibration: Theory and experiments. In *ECCV '92: Proceedings of the Second European Conference on Computer Vision*, pages 321–334. Springer-Verlag, 1992. 22

[58] J. Filip and M. Haindl. Bidirectional texture function modeling: A state of the art survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(11):1921–1940, November 2009. 33

[59] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981. 14

[60] Andrew W. Fitzgibbon and Andrew Zisserman. Automatic camera recovery for closed or open image sequences. In *5th European Conference on Computer Vision*, volume 1406 of *Lecture Notes in Computer Science*, pages 311–326. Springer, 1998. 25

[61] Sing-Choong Foo. A gonioreflectometer for measuring the bidirectional reflectance of material for use in illumination computation. Master's thesis, Program of Computer Graphics, Cornell University, August 1997. xi, 31

[62] J. M. Frahm and M. Pollefeys. RANSAC for (quasi-)degenerate data (QDEGSAC). In *CVPR*, pages I: 453–460, 2006. 14

[63] Thomas Franken, Matteo Dellepiane, Fabio Ganovelli, Paolo Cignoni, Claudio Montani, and Roberto Scopigno. Minimizing user intervention in registering 2d images to 3d models. *The Visual Computer*, 21(8-10):619–628, sep 2005. Special Issues for Pacific Graphics 2005. xi, 15, 53, 65

[64] T Freeth, Y Bitsakis, X Moussas, J H Seiradakis, A Tselikas, H Mangou, M Zafeiropoulou, R Hadland, D Bate, A Ramsey, M Allen, A Crawley, P Hockley, T Malzbender, D Gelb, W Ambrisco, and M G Edmunds. Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism. *Nature*, 444(7119):587–591, 2006. 43

[65] Ran Gal, Yonatan Wexler, Eyal Ofek, Hugues Hoppe, and Daniel Cohen-Or. Seamless montage for texturing models. *Comput. Graph. Forum*, 29(2):479–486, 2010. 46

[66] Andrew Gardner, Chris Tchou, Tim Hawkins, and Paul Debevec. Linear light source reflectometry. *ACM Transactions on Graphics*, 22(3):749–758, July 2003. 37

[67] Pascal Gautron, Jaroslav Křivánek, Sumanta N. Pattanaik, and Kadi Bouatouch. A novel hemispherical basis for accurate and efficient rendering. In *Rendering Techniques 2004, Eurographics Symposium on Rendering*, pages 321–330, June 2004. 80

[68] Todor Georgeiv, Ke Colin Zheng, Brian Curless, David Salesin, Shree Nayar, and Chintan Intwala. Spatio-angular resolution tradeoffs in integral photography. In *Eurographics Workshop/ Symposium on Rendering*, pages 263–272, 2006. 40

[69] A. S. Georghiades. Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In *ICCV*, pages 816–823, 2003. 32

[70] Todor Georgiev and Andrew Lumsdaine. Focused plenoptic camera and rendering. *Journal of Electronic Imaging*, 19(2):021106, 2010. 40

[71] Riccardo Gherardi, Michela Farenzena, and Andrea Fusiello. Improving the efficiency of hierarchical structure-and-motion. In *CVPR*, pages 1594–1600. IEEE, 2010. 25

[72] A. Ghosh, S. Achutha, W. Heidrich, and M. O'Toole. BRDF acquisition with basis illumination. In *ICCV*, pages 1–8, 2007. 31

[73] Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A. Wilson, and Paul Debevec. Circularly polarized spherical illumination reflectometry. *ACM Transactions on Graphics*, 29(6):162:1–162:12, December 2010. 34

[74] Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A. Wilson, and Paul E. Debevec. Estimating specular roughness and anisotropy from second order spherical gradient illumination. *Computer Graphics Forum*, 28(4):1161–1170, 2009. 34

[75] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 129:1–129:10. ACM, 2011. 34

[76] Simon Gibson, Jon Cook, Toby Howard, Roger Hubbold, and Dan Oram. Accurate camera calibration for off-line, video-based augmented reality. In *ISMAR '02: Proceedings of the 1st International Symposium on Mixed and Augmented Reality*, pages 37–46. IEEE Computer Society, 2002. 23

[77] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and spatially-varying BRDFs from photometric stereo. In *ICCV*, pages I: 341–348, 2005. 36

[78] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54. ACM, 1996. 38, 39, 41, 72

[79] Prabath Gunawardane, Oliver Wang, Steven Scher, Ian Rickard, James Davis, and Tom Malzbender. Optimized Image Sampling for View and Light Interpolation. In *The 10th International Symposium on Virtual Reality Archaeology and Cultural Heritage VAST 2009*, pages 93 – 100, 2009. 43

[80] Tom Haber, Christian Fuchs, Philippe Bekaert, Hans-Peter Seidel, Michael Goesele, and Hendrik P. A. Lensch. Relighting objects from image collections. In *CVPR*, pages 627–634. IEEE, 2009. 35

[81] Ø yvind Hammer, Stefan Bengtson, Tom Malzbender, and Dan Gelb. Imaging Fossils using Reflectance Transformation and Interactive Manipulation of Virtual Light Sources. *Paleontologia Electronica*, 2002. 43

[82] K. Hara, K. Nishino, and K. Ikeuchi. Mixture of spherical distributions for single-view relighting. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(1):25–35, January 2008. 32

[83] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Fourth Alvey Vision Conference*, pages 147–151. The Plessey Company plc., 1988. 19, 21

[84] R. I. Hartley. A linear method for reconstruction from lines and points. In *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, page 882. IEEE Computer Society, 1995. 22

[85] Richard I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Computer Vision - ECCV'92, Second European Conference on Computer Vision, Proceedings*, volume 588, pages 579–587. Springer, 1992. 22, 23

[86] Richard I. Hartley. Euclidean reconstruction from uncalibrated views. In *Proceedings of the Second Joint European - US Workshop on Applications of Invariance in Computer Vision*, pages 237–256. Springer-Verlag, 1994. 24

[87] Benno Heigl, Reinhard Koch, Marc Pollefeys, Joachim Denzler, and Luc J. Van Gool. Plenoptic modeling and rendering from image sequences taken by hand-held camera. In *DAGM-Symposium*, pages 94–101, 1999. 41

[88] Janne Heikkila and Olli Silven. A four-step camera calibration procedure with implicit image correction. In *CVPR '97: Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1106–1112. IEEE Computer Society, 1997. 12

[89] Aaron Hertzmann and Steven M. Seitz. Shape and materials by example: A photometric stereo approach. In *CVPR*, pages 533–540. IEEE Computer Society, 2003. 37

[90] A. Heyden and K. Astrom. Euclidean reconstruction from constant intrinsic parameters. In *ICPR '96: Proceedings of the 1996 International Conference on Pattern Recognition*, volume I, pages 339–343. IEEE Computer Society, 1996. 24

[91] Anders Heyden and Kalle Astrom. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, pages 438–443. IEEE Computer Society, 1997. 24

[92] Michael Holroyd, Jason Lawrence, and Todd Zickler. A coaxial optical scanner for synchronous acquisition of 3D geometry and surface reflectance. *ACM Transaction on Graphics*, 29(4), 2010. xi, 35, 36

[93] B. K. P. Horn. Closed form solutions of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America*, 5(7):1127–1135, 1987. 14

[94] K. Ikeuchi, T. Oishi, J. Takamatsu, R. Sagawa, A. Nakazawa, R. Kurazume, K. Nishino, M. Kamakura, and Y. Okamoto. The great buddha project: Digitally archiving, restoring, and analyzing cultural heritage objects. *International Journal of Computer Vision*, 75(1):189–208, October 2007. 15

[95] A. Irschara, C. Zach, and H. Bischof. Towards wiki-based dense city modeling. In *Virtual Representations and Modeling of Large-scale environments*, pages 1–8, 2007. 24

[96] Aaron Isaksen, Leonard McMillan, and Steven J. Gortler. Dynamically reparameterized light fields. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH '00)*, pages 297–306, July 23–28 2000. 41

[97] Jan Jachnik, Richard A. Newcombe, and Andrew J. Davison. Real-time surface light-field capture for augmentation of planar specular surfaces. In *ISMAR*, pages 91–97. IEEE Computer Society, November 2012. 42

[98] H. L. Jin, P. Favaro, and S. Soatto. Real-time feature tracking and outlier rejection with changes in illumination. In *Proceedings of the Eighth International Conference On Computer Vision*, pages I: 684–689, 2001. 22

[99] George Kamberov, Gerda Kamberova, Ondřej Chum, Štěpán Obdržálek, Daniel Martinec, Jana Kostková, Tomáš Pajdla, Jiří Matas, and Radim Šára. 3D geometry from uncalibrated images. In *ISVC '06: Proceedings 2nd International Symposium on Visual Computing*, number 4292 in Lecture Notes in Computer Science, pages 802–813. Springer-Verlag, 2006. 24

[100] Hirokazu Kato and Mark Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *IWAR*, page 85, 1999. 19

[101] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages 506–513, 2004. 20

[102] Khronos Group. WebGL - OpenGL ES 2.0 for the Web. `http://www.khronos.org/webgl/`, 2009. [Accessed on May 2013]. 122

[103] Eric P. F. Lafortune, Sing-Choong Foo, Kenneth E. Torrance, and Donald P. Greenberg. Non-linear approximation of reflectance functions. In *SIGGRAPH 97 Conference Proceedings*, Annual Conference Series, pages 117–126. ACM SIGGRAPH, Addison Wesley, August 1997. 30

[104] Frédéric Larue and Jean-Michel Dischler. Automatic registration and calibration for efficient surface light field acquisition. In *International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, pages 171–178, 2006. 16

[105] Jason Lawrence, Aner Ben-Artzi, Christopher DeCoro, Wojciech Matusik, Hanspeter Pfister, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Inverse

shade trees for non-parametric material representation and editing. *ACM Transactions on Graphics*, 25(3):735–745, July 2006. 33

[106] V. Lempitsky and D. Ivanov. Seamless mosaicing of image-based texture maps. In *CVPR*, pages 1–6, 2007. 46

[107] H. Lensch, W. Heidrich, and H. Seidel. Automated texture registration and stitching for real world models. In *Proceedings of the 8th Pacific Graphics Conference on Computer Graphics and Application (PACIFIC GRAPHICS-00)*, pages 317–327. IEEE, October 3–5 2000. xi, 16, 17, 45

[108] Hendrik P. A. Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics*, 22(2):234–257, April 2003. xi, 34, 35

[109] Stefan Leutenegger, Margarita Chli, and Roland Siegwart. BRISK: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision, ICCV 2011*, pages 2548–2555. IEEE, 2011. 21

[110] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathmatics*, II(2):164–168, 1944. 12

[111] M. E. Leventon, W. M. Wells, and W. E. L. Grimson. Multiple view 2D-3D mutual information registration. In *Image Understanding Workshop*, pages 625–630, 1997. 16

[112] Marc Levoy and Pat Hanrahan. Light field rendering. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42. ACM, 1996. xi, 38, 39, 40, 72

[113] Hongsong Li, Sing Choong Foo, Kenneth E. Torrancei, and Stephen H. Westin. Automated three-axis gonioreflectometer for computer graphics applications. In *Advanced Characterization Techniques for Optics, Semiconductors, and Nanotechnologies II*, volume Proceedings of SPIE Vol. 5878, pages 5878–29. SPIE, SPIE, Bellingham, WA, July 2005. 31

[114] Yunzhen Li and Kok-Lim Low. Automatic registration of color images to 3d geometry. In *CGI '09: Computer Graphics International*, pages 21–28. ACM, 2009. 26

[115] L. Y. Liu and I. Stamos. Automatic 3D to 2D registration for the photorealistic rendering of urban scenes. In *CVPR*, pages II: 137–143, 2005. 15

[116] Lingyun Liu, Ioannis Stamos, Gene Yu, George Wolberg, and Siavash Zokai. Multiview geometry for texture mapping 2d images onto 3d range data. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2293–2300, 2006. xi, 25, 26

[117] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981. 22

[118] M.I.A. Lourakis. levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. `http://www.ics.forth.gr/~lourakis/levmar/`, Jul. 2004. [Accessed on May 2013.]. 53

[119] D. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, 1991. 16

[120] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE Computer Society, 1999. 20

[121] Rong Lu and Jan J. Koenderink. Optical properties (bidirectional reflection distribution functions) of velvet. *Applied Optics*, 37:5974–5984, 1998. 32

[122] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop*, pages 121–130, 1981. 21

[123] Q. T. Luong and T. Vieville. Canonical representations for the geometries of multiple projective views. *Computer Vision and Image Understanding*, 64(2):193–229, September 1996. 24

[124] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the Eurographics Symposium on Rendering Techniques 2007*, pages 183–194. Eurographics Association, 2007. 34

[125] Frederik Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions of Medical Imaging*, 16(2):187–198, April 1997. 16

[126] Frederik Maes, Dirk Vandermeulen, , and Paul Suetens. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Medical Image Analysis*, 3(4):373–386, December 1999. 16

[127] Tom Malzbender, Dan Gelb, and Hans Wolters. Polynomial texture maps. In *SIGGRAPH 2001, Computer Graphics Proceedings*, Annual Conference Series, pages 519–528. ACM Press / ACM SIGGRAPH, 2001. 36, 43, 111

[128] Stephen R. Marschner, Stephen H. Westin, Eric P. F. Lafortune, and Kenneth E. Torrance. Image-based bidirectional reflectance distribution function measurement. *Applied Optics*, 39(16):2592–2600, June 2000. 32

[129] Kenji Matsushita and Toyohisa Kaneko. Efficient and handy texture mapping on 3D surfaces. *Computure Graphics Forum*, 18(3):349–358, 1999. 15

[130] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. In *Proceedings of ACM SIGGRAPH*, volume 22(3) of *ACM Transactions on Graphics*, pages 759–769, 2003. 32

[131] Wojciech Matusik, Hanspeter Pfister, Matthew Brand, and Leonard McMillan. Efficient isotropic BRDF measurement. In *Proceedings of the 14th Eurographics workshop on Rendering*, pages 241–248. Eurographics Association, 2003. 32

[132] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, October 2005. 21

[133] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Proceedings of the Eighth International Conference On Computer Vision*, pages 525–531, 2001. 20

[134] Gavin Miller, Steven Rubin, and Dulce Ponceleon. Lazy decompression of surface light fields for precomputed global illumination. In *Rendering Techniques '98*, Eurographics, pages 281–292. Springer-Verlag Wien New York, 1998. 39, 72, 74

[135] Mark Mudge, Tom Malzbender, Alan Chalmers, Roberto Scopigno, James Davis, Oliver Wang, Prabath Gunawardane, Michael Ashley, Martin Doerr, Alberto Proenca, and João Barbosa. Image-Based Empirical Information Acquisition , Scientific Reliability , and Long-Term Digital Preservation for the Natural Sciences and Cultural Heritage. In *Tutorial Eurographics 08*. Eurographics, 2008. 126

[136] Mark Mudge, Tom Malzbender, Carla Schroer, and Marlin Lum. New Reflection Transformation Imaging Methods for Rock Art and Multiple-Viewpoint Display. In *The 7th International Symposium on Virtual Reality Archaeology and Cultural Heritage VAST 2006*, pages 195–202, 2006. 43

[137] Mark Mudge, Jean-Pierre Voutaz, Carla Schroer, and Marlin Lum. Reflection Transformation Imaging and Virtual Representations of Coins from the Hospice of the Grand St. Bernard. In *The 6th International Symposium on Virtual Reality Archaeology and Cultural Heritage VAST 2005*, pages 29–39. Eurographics Association, 2005. 43

[138] Gero Müller, Jan Meseth, Mirko Sattler, Ralf Sarlette, and Reinhard Klein. Acquisition, synthesis and rendering of bidirectional texture functions. In *Eurographics 2004, State of the Art Reports*, pages 69–94, 2004. 27, 33

[139] Takeshi Naemura, Junji Tago, and Hiroshi Harashima. Real-time video-based modeling and rendering of 3D scenes. *IEEE Computer Graphics and Applications*, 22(2):66–73, March/April 2002. xi, 40, 41

[140] Leonid Naimark and Eric Foxlin. Encoded LED system for optical trackers. In *ISMAR*, pages 150–153, 2005. 19

[141] Peter J. Neugebauer and Konrad Klein. Texturing 3D Models of Real World Objects from Multiple Unregistered Photographic Views. *Computer Graphics Forum*, 18(3):245–256, September 1999. 15, 45

[142] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a Hand-Held plenoptic camera. Technical report, Stanford University, April 2005. 40

[143] Addy Ngan, Frédo Durand, and Wojciech Matusik. Experimental analysis of brdf models. In *Proceedings of the Eurographics Symposium on Rendering*, pages 117–226. Eurographics Association, 2005. 32

[144] K. Ni, D. Steedly, and F. Dellaert. Out-of-core bundle adjustment for large-scale 3D reconstruction. In *ICCV*, pages 1–8, 2007. 25

[145] F. E. Nicodemus, J. C. Richmond, J. J. Hsia, I. W. Ginsberg, and T. Limperis. Geometric considerations and nomenclature for reflectance. Monograph 161, National Bureau of Standards (US), October 1977. 28

[146] Ko Nishino, Zhengyou Zhang, and Katsushi Ikeuchi. Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. In *ICCV*, pages 599–606, 2001. 35

[147] F. Moreno Noguer, V. Lepetit, and P. Fua. Accurate non-iterative O(n) solution to the pnP problem. In *Computer Vision, IEEE International Conference on*, pages 1–8, 2007. 14

[148] D. Oberkampf, D. F. DeMenthon, and L. S. Davis. Iterative pose estimation using coplanar feature points. *Computer Vision and Image Understanding*, 63(3):495–511, May 1996. 13

[149] Michael Oren and Shree K. Nayar. Generalization of lambert's reflectance model. In *SIGGRAPH '94: Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 239–246. ACM, 1994. 30

[150] Joseph Padfield, David Saunders, and Tom Malzbender. Polynomial texture mapping: a new tool for examining the surface of paintings. *ICOM Committee for Conservation*, I:504–510, 2005. 43

[151] G. Panin and A. Knoll. Mutual information-based 3D object tracking. *International Journal of Computer Vision*, 78(1):107–118, June 2008. 18

[152] Matt Pharr and Simon Green. Ambient occlusion. In Randima Fernando, editor, *GPU Gems*, chapter 17, pages 279–292. Addison-Wesley, 2004. 57

[153] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 30

[154] Nico Pietroni, Marco Tarini, and Paolo Cignoni. Almost isometric mesh parameterization through abstract domains. *IEEE TVCG*, 16(4):621–635, July/August 2010. 101

[155] Ruggero Pintus, Enrico Gobbetti, and Roberto Combet. Fast and robust semi-automatic registration of photographs to 3D geometry. In *The 12th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*, pages 9–16, October 2011. 26

[156] Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever. Mutual information based registration of medical images: A survey. *IEEE Transaction on Medical Imaging*, 22(8):986–1004, 2003. 16

[157] M. Pollefeys, L. J. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, and J. Tops. Video-to-3D. In *Photogrammetric Computer Vision*, page A: 252, 2002. 23

[158] M. Pollefeys, R. Koch, and L. J. Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32(1):7–25, August 1999. 24

[159] Damien Porquet, Jean-Michel Dischler, and Djamchid Ghazanfarpour. Real-time high-quality view-dependent texture mapping using per-pixel visibility. In *Proceedings of the 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, pages 213–220, 2005. 46

[160] M. J. D. Powell. Developments of NEWUOA for minimization without derivatives. *IMA Journal of Numerical Analysis*, 28(4):649–664, October 2008. 18, 58

[161] V S Ramachandran. Perception of shape from shading. *Nature*, 331(6152):163–166, 1988. 124

[162] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *SIGGRAPH 2001 Conference Proceedings*, pages 117–128. ACM Press, 2001. 33

[163] F. Remondino and C. Fraser. Digital camera calibration methods: considerations and comparisons. In Isprs, editor, *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,*, volume Vol. XXXVI, 2006. 13

[164] Peiran Ren, Jiaping Wang, John Snyder, Xin Tong, and Baining Guo. Pocket reflectometry. *ACM Transaction on Graphics*, 30(4):45, 2011. xi, 37

[165] Miguel Ribo and Axel Pinz. A new optical tracking system for virtual and augmented reality applications. In *IEEE Instrumentation and Measurement Technical Conference*, pages 1932–1936, 2001. 19

[166] Claudio Rocchini, Paolo Cignomi, Claudio Montani, and Roberto Scopigno. Multiple textures stitching and blending on 3D objects. In *Rendering Techniques '99*, Eurographics, pages 119–130. Springer-Verlag Wien New York, 1999. 45

[167] Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837, September 1956. 100

[168] E. Rosten and T. W. Drummond. Machine learning for high-speed corner detection. In *ECCV*, pages I: 430–443, 2006. 20

[169] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: An efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision, ICCV 2011*, pages 2564–2571. IEEE, 2011. 21

[170] Szymon Rusinkiewicz, Michael Burns, and Doug DeCarlo. Exaggerated shading for depicting shape and detail. *ACM Transactions on Graphics*, 25(3):1199–1205, jul 2006. 44

[171] Szymon M. Rusinkiewicz. New change of variables for efficient BRDF representation. In *Rendering Techniques '98*, Eurographics, pages 11–22. Springer-Verlag Wien New York, 1998. 29

[172] J. Salvi, X. Armangue, and J. Batlle. A comparative review of camera calibrating methods with accuracy evaluation. *Pattern Recognition*, 35(7):1617–1635, July 2002. 13

[173] H. Schirmacher, L. Ming, and H.-P. Seidel. On-the-Fly processing of generalized lumigraphs. In *Proceedings of the 22nd Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS-01)*, volume 20, 3 of *Computer Graphics Forum*, pages 165–173, September 4–7 2001. 40

[174] Christopher Schwartz, Michael Weinmann, Roland Ruiters, and Reinhard Klein. Integrated high-quality acquisition of geometry and appearance for cultural heritage. In *VAST 2011*, pages 25–32, October 2011. 34

[175] S. A. Shafer. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985. 73, 74

[176] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, June 1994. 21

[177] H.-Y. Shum, Q. Ke, and Z. Zhang. Efficient bundle adjustment with virtual key frames: A hierarchical approach to multi-frame structure from motion. In *Proceedings of the IEEE Computer Science Conference on Computer Vision and Pattern Recognition (CVPR-99)*, pages 538–543. IEEE, June 23–25 1999. 25

[178] Heung-Yeung Shum and Li-Wei He. Rendering with concentric mosaics. In *Proceedings of the Conference on Computer Graphics (SIGGRAPH'99)*, pages 299–306, August8–13 1999. 40

[179] S. M. Smith and J. M. Brady. SUSAN - a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, May 1997. 20

[180] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph*, 25(3):835–846, 2006. 25

[181] I. Stamos, L. Y. Liu, C. Chen, G. Wolberg, G. Yu, and S. Zokai. Integrating automated range registration with multiview geometry for the photorealistic modeling of large-scale scenes. *International Journal of Computer Vision*, 78(2-3):237–260, July 2008. 25

[182] D. Steedly, I. A. Essa, and F. Dellaert. Spectral partitioning for structure from motion. In *ICCV*, pages 996–1003, 2003. 25

[183] P. F. Sturm and R. I. Hartley. Triangulation. In *Image Understanding Workshop*, pages II:957–966, 1994. 24

[184] Thorsten Thormaehlen and Hellward Broszio. Voodoo Camera Tracker. `http://www.digilab.uni-hannover.de/docs/manual.html`, 2002. [Accessed on May 2013.]. 53

[185] T. Thormahlen, H. Broszio, and A. Weissenfeld. Keyframe selection for camera motion and structure estimation from multiple views. In *Proceedings of the 8th European Conference on Computer Vision*, pages Vol I: 523–535, 2004. 23

[186] Thorsten Thormählen, Nils Hasler, Michael Wand, and Hans-Peter Seidel. Merging of feature tracks for camera motion estimation from video. In *CVMP*, 2008. 22

[187] Corey Toler-Franklin, Adam Finkelstein, and Szymon Rusinkiewicz. Illustration of complex real-world objects using images with normals. In *Proceedings of the 5th International Symposium on Non-Photorealistic Animation and Rendering 2007*, pages 111–119. ACM, 2007. 44

[188] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, Computer Science Department, Carnegie Mellon University, April 1991. 21

[189] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, April 2000. 14

[190] A. Treuille, A. Hertzmann, and S. M. Seitz. Example-based stereo with general BRDFs. In *ECCV*, pages Vol II: 457–469, 2004. 37

[191] B. Triggs, P. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment: A modern synthesis. In *Vision Algorithms Workshop: Theory and Practice*, 1999. 24

[192] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, 3:323–344, 1987. 12, 51, 65

[193] T. Tuytelaars and L. J. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proceedings of the British Machine Vision Conference*, pages 412–425, 2000. 20

[194] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2007. 20

[195] A. Veeraraghavan, A. Agrawal, R. Raskar, A. Mohan, and J. Tumblin. Non-refractive modulators for encoding and capturing scene appearance and depth. In *CVPR*, pages 1–8, 2008. 40

[196] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: mask enhanced cameras for heterodyned

light fields and coded aperture refocusing. *ACM Transactions on Graphics*, 26(3):69:1–69:11, July 2007. 40

[197] M. Vergauwen and L. J. Van Gool. Web-based 3D reconstruction service. *Machine Vision and Applications*, 17(6):411–426, December 2006. 24

[198] Romain Vergne, Romain Pacanowski, Pascal Barla, Xavier Granier, and Christophe Schlick. Light warping for enhanced surface depiction. *ACM Transactions on Graphics*, 28(3), 2009. 44

[199] P. A. Viola and W. M. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, September 1997. xi, 16, 17

[200] Jiaping Wang, Shuang Zhao, Xin Tong, John Snyder, and Baining Guo. Modeling anisotropic surface reflectance with example-based microfacet synthesis. *ACM Transaction on Graphics*, 27(3), 2008. 37

[201] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transaction on Image Processing*, 13(4):600–612, April 2004. 83

[202] Gregory J. Ward. Measuring and modeling anisotropic reflection. In *Computer Graphics (SIGGRAPH '92 Proceedings)*, volume 26, pages 265–272, July 1992. 30, 31

[203] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics*, 24(3):756–764, July 2005. 34

[204] D. Rod White, Peter Saunders, Stuart J. Bonsey, John van de Ven, and Hamish Edgar. Reflectometer for measuring the bidirectional reflectance of rough surfaces. *Applied Optics*, 37(16):3450–3454, Jun 1998. 31

[205] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. *ACM Transactions on Graphics*, 24(3):765–776, July 2005. xi, 40, 41

[206] Geert Willems, Frank Verbiest, Wim Moreau, Hendrik Hameeuw, Karel Van Lerberghe, and Luc Van Gool. Easy and cost-effective cuneiform digitizing. In *The 6th International Symposium on Virtual Reality Archaeology and Cultural Heritage VAST 2005*, pages 73–80. Eurographics Association, 2005. 43, 126

[207] Daniel N. Wood, Daniel I. Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, Werner Stuetzle, and David H. Salesin. Surface light fields for 3D photography. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH'00)*, pages 287–296, July 23–28 2000. xi, 42

[208] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *ICCV*, pages 1108–1115, 2011. 36

[209] Chenglei Wu, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *CVPR*, pages 969–976, 2011. 36

[210] Hongzhi Wu, Julie Dorsey, and Holly E. Rushmeier. A sparse parametric mixture model for BTF compression, editing and rendering. *Computer Graphics Forum*, 30(2):465–473, 2011. 33

[211] Jason C. Yang, Matthew Everett, Chris Buehler, and Leonard McMillan. A real-time distributed light field camera. In *13th Eurographics Workshop on Rendering*, pages 77–86, 2002. 40

[212] Tianli Yu, Hongcheng Wang, Narendra Ahuja, and Wei-Chao Chen. Sparse lumigraph relighting by illumination and reflectance estimation from multi-view images. In *Eurographics Workshop/ Symposium on Rendering*, pages 41–50. Eurographics Association, 2006. 36

[213] Cha Zhang and Tsuhan Chen. A self-reconfigurable camera array. In *Eurographics Symposium on Rendering*, pages 243–254, 2004. 40

[214] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 13

[215] W. Zhao, D. Nister, and S. Hsu. Alignment of continuous video onto 3D point clouds. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(8):1305–1318, August 2005. 25

[216] Hongwei Zheng, Ioan Cleju, and Dietmar Saupe. Highly-automatic MI based multiple 2D/3D image registration using self-initialized geodesic feature correspondences. In *ACCV*, volume 5996 of *Lecture Notes in Computer Science*, pages 426–435. Springer, 2009. 26

[217] Todd Zickler, Sebastian Enrique, Ravi Ramamoorthi, and Peter Belhumeur. Reflectance sharing: Image-based rendering from a sparse set of images. In *Eurographics Symposium on Rendering*, pages 253–264. Eurographics Association, 2005. 34