A Survey and Tutorial on Shape and Reflectance Estimation Techniques

Dana Cobzas\textsuperscript{1}, Neil Birkbeck\textsuperscript{1}, Martin Jagersand\textsuperscript{1} and Peter Sturm\textsuperscript{2}

\textsuperscript{1}Computer Science, University of Alberta, Canada
\textsuperscript{2}INRIA Rhone-Alpes, France
Abstract

Recently there has been significant activity and progress in reconstruction of surface and reflectance models from images of objects with a general reflectance. Some recent work builds-on, extends and combines multi-view stereo, volumetric reconstruction and various shape-from-shading approaches. Another characteristic of the more recent work is a that a careful mathematical modeling separates the problem formulation (e.g. cost functional) from an algorithmic formulation (PDE solver, optimization) as in the variational approaches to shape and reflectance recovery. In this paper we combine a survey of recent methods with a gentle enough introduction/tutorial to make it accessible to a wide audience. In particular we are provide a taxonomy of methods, grouping them mainly according to the type of representation (disparity map, voxels, mesh, level sets). We present state of the art optimization approaches used in surface reconstruction like graph cuts or variational methods. Special attention is given to methods that deal with general reflectance for both single view and multi-view approaches. To make the paper readable standalone, we provide short reviews of important theoretical aspects of the shape reconstruction (e.g. reflectance modeling, shape from shading, photometric stereo, variational stereo formulation).

Index Terms

shape reconstruction, surface reflectance reconstruction, survey, shape from shading, photometric stereo, volumetric reconstruction, variational, level sets, multi-view stereo

I. INTRODUCTION

The automatic computation of 3D geometric and appearance models from images is one of the most challenging and fundamental problems in computer vision. The focus of this paper is to provide a survey of existing techniques in surface and reflectance reconstruction.

Most studied approaches in geometric reconstruction, called structure-from-motion (SFM), use the 2D position of physical points or other features in several images, called corresponding features, to recover their 3D coordinates. Recent research in non-Euclidean geometry [56] has proved that it is possible to recover both structure (3D feature positions) and camera positions from corresponding points in uncalibrated images. The result is a collection of 3D features, that characterize the scene but without any information on the actual shape (or surface). During recent years methods that attempt to reconstruct a detailed shape/surface model for an object or scene became increasingly popular. Using the notion of an object surface, information about light and
reflectance or material properties can be incorporated in the formulation, making the problem more general and eliminating the usual assumption of Lambertian scenes. Besides, the continuity and smoothness of the surface can be imposed through regularization. This paper provides a survey of existing techniques for reconstructing shape and reflectance from multiple images. Our goal is to present a taxonomy of the current approaches and also to discuss and contrast state of the art optimization methods for solving the reconstruction problem (e.g. graph-cuts and variational methods). We give special attention to methods that deal with specular reflectance and explicitly consider shading cues in the reconstruction. These methods are less common than the traditional ones based on stereo cues. For completeness we shortly review important theoretical aspects in reconstruction giving the paper a tutorial part and making the survey easier to understand for people that are new in the field. To summarize, the main contributions of our survey are the following:

- We present a survey and taxonomy of methods that reconstruct surface (dense shape) from images mainly grouped by the type of representation (disparity map, voxels, mesh, level sets). The survey deals only with static scenes.
- We present the current state-of-the-art techniques for methods designed for objects with general reflectance for both single view and multiple view approaches.
- The tutorial part of the survey concisely presents basic theoretical aspects involved in the reconstruction problem (e.g. reflectance modeling, shape from shading, photometric stereo, variational stereo formulation). This part makes the survey accessible to a larger group of readers.

As the general shape reconstruction problem is vast – ranging from methods for shape-from-shading and photometric stereo to volumetric reconstruction and variational multi-view stereo – it is beyond the scope of a paper to provide a complete study, so we give an overall understanding of different subproblems, (i) to identify classes of techniques and (ii) to collect major theoretical results beyond the existing surveys of some subclasses (shape from shading [164], volumetric reconstruction methods [133], and two view stereo [123]).

Methods of shape reconstruction involve recovering the shape (and reflectance) of a scene or object from one or more images taken under the same or different illumination conditions. We will first identify the important characteristics or dimensions that will be used for the classification. We
divide them in several categories. The first categories (number of images, light, and reflectance) are connected with the data and assumptions about it while the later ones (surface representation, optimization, and information used) are connected with the method used in recovering the shape (and reflectance). Table I presents a summary and a detailed presentation follows.

<table>
<thead>
<tr>
<th>No. images/views</th>
<th>Light conditions</th>
<th>Reflectance function</th>
</tr>
</thead>
<tbody>
<tr>
<td>single image</td>
<td>single source constant known</td>
<td>non-Lambertian uniform param. model</td>
</tr>
<tr>
<td>multi images, single view</td>
<td>finite no. sources varying accounted</td>
<td>[interrelations] non-param. model</td>
</tr>
<tr>
<td>multi views</td>
<td>general light recovered</td>
<td>[opacity] accounted</td>
</tr>
<tr>
<td>[weak calib]</td>
<td>[distant/finite]</td>
<td></td>
</tr>
<tr>
<td>[dynamic]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image information</th>
<th>Surface representation</th>
<th>Optimization method</th>
</tr>
</thead>
<tbody>
<tr>
<td>silhouette</td>
<td>depth/disparity map</td>
<td>linear, non-linear</td>
</tr>
<tr>
<td>texture</td>
<td>3D mesh</td>
<td>variational</td>
</tr>
<tr>
<td>shading</td>
<td>volumetric grid</td>
<td>graph cuts</td>
</tr>
<tr>
<td>specular highlights</td>
<td>level sets</td>
<td>simulated annealing, mean-field annealing</td>
</tr>
<tr>
<td>shadows</td>
<td></td>
<td>belief propagation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dynamic programming [...]</td>
</tr>
</tbody>
</table>

TABLE I

DIMENSIONS FOR CLASSIFICATIONS AND THE MOST REPRESENTATIVE CATEGORIES FOR THE PROBLEM OF SHAPE ESTIMATION

A. Images/views

Shape can be recovered using a single image (e.g. shape from shading) or multiple images that represent the same viewpoint (e.g. photometric stereo) or multiple viewpoints of the scene (e.g. multi-view stereo). In the case of multiple views we assume that the images have been calibrated, meaning that the position of the camera with respect to the object is known for each view. There exist some approaches that work with only weakly calibrated cameras like the volumetric methods presented by Kimura et. al [73] and Saito et. al [117]. Also, the variational approach presented by Alvarez and Deriche[1] recovers a projective depth of the scene. In the case of one or multiple images taken from the same view there is no need for calibration as the correspondence problem is implicitly solved. A depth and normal for each image pixel is recovered in this case.

April 26, 2006
B. Light

There are two different aspects to consider about light: what type of light is illuminating the scene (light parameters) and what extent it is recovered. Regarding the type of light it can be a single light source, a finite number of light sources or a general light configuration that can for example be represented as a discretization of the illumination hemisphere or with spherical harmonic coefficients [114]. When the lights are modeled explicitly with point light sources another consideration is whether they are distant (infinite) or finite. Most of the work in the literature consider distant light sources. For the finite light case, we can mention an important result of Prados and Faugeras [110] who showed that if the attenuation factor is considered, the shape from shading problem becomes well-posed.

Regarding the knowledge about the light conditions we identified four cases – known when the position and strength of the light is assumed known; accounted when the reconstruction method accounts for the light change through a robust photo-consistency measurement (e.g. the normalized cross correlation score accounts for light variations caused by a Lambertian object moving with respect to the light); and recovered when the parameters of the light are explicitly recovered together with the shape of the scene.

It is worth making a note about the types of setup for taking the images considering light variation with respect to the scene. Here are the main cases we identified in the literature: I) multiple images from the same view of a scene with changing light – photometric stereo setup when when the light is explicitly changed and the camera does not move; II) multiple views of the object under the same light condition – traditional multi-view stereo case when the camera moves with respect to the object; III) multiple views with different illumination conditions that are obtained by moving the object with respect to the scene keeping the camera fixed. The last two multi-view cases are equivalent for the structure from motion problem but when considering explicit reflectance modeling they are different as the light is stationary with respect to the scene in the first case while it is moving in the second case. So we can add another characteristic about light whether it is constant or variable.

C. Reflectance

The surface appearance change when the light and viewpoint are changing is modeled by the reflectance function. The material properties that generate the change can be represented using
the BRDF (bi-directional reflectance distribution function). Section II-A gives a short summary of the reflectance theory. Like in the case of light we considered two aspects, one connected to the type of reflectance function and the other with its representation and method used in recovery. We chose to classify different types of reflectance in: Lambertian/non-Lambertian, textured/non-textured. We refer to objects with uniform color as non-textured and to objects with varying colors as textured. A Lambertian BRDF is the usual approximation for the diffuse component. There are three main approaches to deal with specular (non-Lambertian) reflections. Specular parts can be filtered/separated from the diffuse ones and the reconstruction algorithm is applied only for the diffuse parts. We can place here also the methods that account for non-Lambertian reflectance through special camera-light configurations (e.g. Zickler and Belhumeur [92], [168] use Helmholtz reciprocity infer shape for objects with arbitrary reflections). Other approaches in dealing with general reflectance are to approximate the BRDF with either a parametric model (e.g. Phong, Ward, Torrace-Sparrow) or a non-parametric model. A non-parametric BRDF can be measured or calculated as a 4D map (ignoring wavelength). Under certain scene constraints the dimensionality can be reduced (e.g. the 2D view independent map that encodes the specular BRDF [155] in the case when the light is constant with respect to the scene; or the Hertzmann and Seitz’s shape by examples method [58] that uses a known sphere from the same material as the object as reference for reflectance properties).

When modeling surface properties most of the methods do not consider inter-reflections. We can mention as an exception the work of Chandraker et al. [20] that shows that the bas relief ambiguity is completely resolved for general 3D objects. Also with the exception of DeBonet and Viola [9], and Szeliski and Golland [142] the techniques assume that the surfaces are opaque.

D. Surface representation

This characteristic represents the way the continuous surface is discretized. Some methods (e.g. variational methods, level sets methods) are derived for a continuous surface and only at the time of implementation there is need for discretization. Others like the volumetric methods are defined directly on the discretized space. The most common surface representations are: depth/disparity with respect to one reference view, all the views or a base mesh; 3D triangulated mesh; 3D voxel grid; and level sets. Some methods use two representations.
E. Image information used

There is different types of information that can be extracted from images, referred as image cues. One type of information is related to the silhouette (contour) of the object and the others to the color of different pixels. From color different types of information that can be extracted: texture, shading, specular highlights, cast or attached shadows. While shading is a useful information even for a single image, texture is most often used as stereo cues that relate pairs of images.

Most methods ignore global illumination effects. Some exceptions exist like Lager and Zucker’s work on shape from shading on a cloudy day [81] or Kriegman’s et al. [20] paper on uncalibrated photometric stereo with interreflections.

F. Optimization method

The shape recovery problem is often posed as finding the surface that minimizes a global cost functional. The cost functional involves all the constraints that can be derived on the surface using the image information and assumptions about the scene and light. The resulting minimization problem is then solved using either continuous or discrete methods. The method is often directly dependent on the surface representation. For example, with the voxel based representation, the optimization is purely algorithmic. Other discrete methods are applied for the multi-view stereo reconstruction formulated as a labeling problem. This leads to a NP hard problem that can however be solved using Markov Random Field (MRF) minimization methods like graph cuts, simulated annealing, EM mean-field annealing, belief propagation, dynamic programing. Among these, graph cuts algorithms are quite efficient and proved to be able to find a global solution (e.g. [103]) for a convex interaction (smoothing) energy term or a strong local minima for a large class of energies including some that model discontinuities [13]. Loopy belief propagation is also an efficient adaptive algorithm [140] defined on a graph based on message passing between nodes. But, for graphs containing cycles, is not guaranteed to converge. Continuous methods are mainly based on the variational formulation of the reconstruction problem either using an explicit (e.g. depth map) or an implicit level set representation. Those methods provide an elegant formulation but will only find a local minimum so they are quite sensitive to initialization.

Most of the methods assume that the scene is static while the camera is taking pictures. There exist approaches that reconstruct dynamic scenes (as voxels or depth maps) from images taken by
several cameras placed in the scene. In this case the motion/scene fbw is reconstructed together with the scene [165], [146], [18], [162]. Due to space limitations we limit this review to method that reconstruct static scenes.

We present here methods that reconstruct shape and possible light and reflectance from images and do not consider methods that assume known shape and reconstruct only light (inverse light problem) [94] or reflectance (reflectance modeling) [61], [157] or both (radiometric reconstruction) [91], [99], [55], [121], [114].

All the above criteria can possibly provide a principal dimension for classification. In some cases there is no clean division between different categories and the reconstruction method can depend either on the number of images and the assumptions about light (e.g. shape from shading, photometric stereo) or, in some other cases, it depends on surface representation (e.g. volumetric reconstruction). Therefore, it is difficult to choose only one dimension for guiding the classification so we identified different categories of methods depending on number of views, surface representation and optimization method. Different light conditions and reflectance functions are discussed for every method. We start with a general short introduction of the camera and reflectance models in Section II. We begin the review with the shape-from-shading and photometric stereo techniques in Sections III and IV respectively. We then present in Section V existing approaches for multi-view methods. We start with multi-view stereo, then introduce the volumetric reconstruction methods and mesh-based techniques. We next present one important classes of discrete methods, graph-cuts in Subsection V-D and the continuous approaches (variational and level sets) in Subsection V-E. At the end we discuss different approaches in dealing with objects with general reflectance.

II. Generalities and notations

An image $I$ of a scene is a result of both geometric and photometric properties of the scene (including light conditions). We refer to a pixel (image location) as $x = [u, v]^T$ and to the color or intensity of the image in that location as $I(x) = I(u, v)$ which is a scalar for intensity images or a 3D vector containing image values for the three color channels in the case of color images. A 3D point of the scene is denoted $X = [X, Y, Z]^T$. We use the same notations $x, X$ for the homogeneous coordinates of 2D image points and 3D scene points.

The geometric part of image formation gives the correspondence between a point in 3D (or
more precisely a 3D ray) and an image point. The most common model is a 3D-2D transformation that projects scene points onto image pixels. We denote the $3 \times 4$ projection matrix transformation by $P$ so the image projection can be written in homogeneous coordinates as $x = PX$. For the pinhole (projective) camera model $P$ has the form $P = K[R|t]$ where $K$ is the $3 \times 3$ camera matrix containing the internal camera parameters, and $[R|t]$ is a 3D transformation that aligns the reference coordinate system of the scene with the camera coordinate system (external parameters). Most of the presented shape reconstruction approaches assume that the camera is calibrated meaning that $P$ is known.

A. Reflectance models

The color of an image pixel is determined by the interaction of light with the surface location corresponding to that pixel and by the location of the camera with respect to the scene. The process is more precisely described next.

The *radiance* is the power traveling at some point in a specific direction per unit area perpendicular to the direction of travel, per solid angle. We denote it by $l^\lambda(\omega)$ where the angle $\omega = (\theta, \phi)$ is defined like in Figure 1 (left) and $\lambda$ is the wavelength of light. When the light is not directional (infinite) but finite there is an extra fall-off factor contained in $l^\lambda(\omega)$ (usually modeled

![Diagram](image-url)
as proportional with the inverse of the square distance between the light and surface point). The irradiance represents the unit light (energy) arriving at a surface point \( X \). The irradiance from a direction \( \omega \) is

\[
E_i(\omega) = l(\phi, \theta) \cos(\theta) d\omega
\]

(1)

Integrating it over the whole illumination hemisphere gives the total incoming energy for a surface point

\[
E_i = \int_\Omega l(\phi, \theta) \cos(\theta) \sin(\theta) d\theta d\phi
\]

(2)

Images give a measurement of the energy arriving at the camera or the outgoing radiance \( E_o \) that depends on the surface material properties and the way it reflects light. This light-surface interaction is modeled with the bi-directional reflectance distribution function (BRDF). The BRDF \( \rho^\lambda(\theta_i, \phi_i, \theta_o, \phi_o) \) measures, for a given wavelength, the fraction of incoming irradiance from a direction \( \omega_i = (\theta_i, \phi_i) \) in the outgoing direction \( \omega_o = (\theta_o, \phi_o) \) (see Figure 1 right). The measured radiance can then be expressed in a reflectional equation

\[
E_o = \int_\Omega \rho^\lambda(\theta_i, \phi_i, \theta_o, \phi_o) l(\phi_i, \theta_i) \cos(\theta_i) d\omega_i
\]

(3)

A general BRDF can be expressed as a discrete 4D map (ignoring \( \lambda \)) if sufficient sample data is available for different illumination and viewing conditions [157], [98]. An alternative is to model the BRDF with a parametric model. The most simple model, named Lambertian, models the BRDF with a constant \( \rho^\lambda(\theta_i, \phi_i, \theta_o, \phi_o) = k_d^\lambda \). As a consequence, the light is reflected equally in all directions so the emitted radiance does not depend on view direction. It represents a good model for perfectly diffuse materials like clay. General non-diffuse reflectance is usually modeled as a sum of a diffuse Lambertian component and a specular component. A simple model for specular materials is the Phong model [106]:

\[
\rho^\lambda(\theta_i, \phi_i, \theta_o, \phi_o) = k_d^\lambda + k_s^\lambda (\frac{\cos\theta_r}{\cos\theta_i})^n
\]

(4)

where \( \theta_r \) is the angle between the reflection of the light direction (with respect to the normal) and the view direction (see Figure 1), \( k_s^\lambda \) is the specular color and \( n \) a specular coefficient modeling the sharpness of the specularity. The Phong model does not obey natural properties of a surface [84] (energy conservation, Helmholtz reciprocity) but because of its simplicity is
widely used in computer graphics. Better models include physically-based models (e.g. Torrace-Sparrow [144], Lafortune [79] for specular; Nayar [95] for an extended Lambertian) or phenomenological models [74]. As an example the Torrace-Sparrow model has the form

\[
\rho^\lambda(\theta_i, \phi_i, \theta_o, \phi_o) = k_d^\lambda + k_s^\lambda \frac{Q F}{\cos \theta_o} \exp \left( -\frac{\theta_o^2}{\sigma^2} \right) \tag{5}
\]

\(\sigma\) being the specular roughness coefficient, \(G\) is the geometric attenuation factor (bistatic shading) and \(F\) is the Fresnel reflectance coefficient.

The actual colors returned by the camera might not be an accurate measurement for the outgoing radiance due to the non-linear response function of the camera. But most of the methods assume it is close enough. Accurate measurement to recover the nonlinear mapping were performed using multiple images with different aperture/exposure [28], [96] or illumination [70] or by utilizing an image of a known calibration color pattern [21].

III. SHAPE FROM SHADING

Since shape-from-shading (SFS) was first defined by Horn in the 70s [60] it became a classical computer vision problem. The SFS problem is posed as recovering shape (per-pixel surface orientation) from a single image using information about shading. The classical formulation assumes: infinite light source, orthographic camera, known Lambertian reflectance, and a smooth (differentiable) surface. The surface is assumed to be uniform (unit albedo) so its shading is induced only by the light direction. The normals are then integrated to reconstruct the per-pixel depth. This formulation is ill-posed as there is no unique solution for surface orientation and additional constraints have been defined to further constrain the solution.

A good survey paper by Zhang et. al. [164] reviews and compares results for the most important approaches of the traditional SFS techniques. They classify the existing techniques into four groups based on the type of method used in recovering the normals: minimization approaches (non-linear minimization of an energy function involving additional constraints like the brightness, integrability [46] or intensity gradient constraints [166]), propagation approaches (propagate the shape information from a set of known surface points (e.g. singular points) [60], [72]), local approaches (particular surface types e.g. locally spherical [105]) and linear approaches (linearization of the reflectance map [104]). The study found that none of the algorithms has consistent performance for all test images (synthetic and real). This is due to the inherent
difficulties from the original formulation of the SFS problem that was proved to be *ill-posed* [16], [34]. As a consequence SFS techniques result in *non-convergent algorithms* as pointed out by Durou and Maitre [33].

For a better understanding of the SFS problem we next present its continuous formulation, the nature of resulting PDEs, their convergence and numerical solutions following the work of Prados et al [37], [111], [112]. Their work studies ways of reformulating the SFS problem as a well-posed problem and to increase the convergence of the existing algorithms. Direct numerical solutions for the SFS equations are propagation methods (a recent review was presented by Durou et al [32]). They have some advantages over the other mentioned classes as they do not make any linearization and they do not introduce any bias (unlike nonlinear methods that add regularization or integrability constraints). Efficient numerical methods based in the level set formulation of the problem have been developed by Kimmel and Sethian [126] for the orthographic case or by Tankus et al [143] for the perspective case.

Remember the measured radiance for a Lambertian object at a given point $X$ illuminated by a directional light source $L$, assuming the color of light and object albedo are 1:

$$E(X) = \cos(n(X), L) = \frac{n(X) \cdot L}{|n(X)|}$$

where $n(X)$ represents the surface normal at point $X$. If a surface is represented as a depth map with respect to the image, in the case of *orthographic* projection the parametrization becomes:

$$S = \{(x, f(x))| x = (u, v) \in \Omega\}$$

$$n(S(x)) = (\nabla f(x), -1)$$

$\Omega$ being the image domain. The resulting reflectance equation is:

$$I(x) = \frac{-\nabla f(x) \cdot 1 + c}{\sqrt{1 + |\nabla f(x)|^2}} \quad (6)$$

where $L = (l, c)$. In the *perspective* camera case with focal length $\lambda$ the surface parametrization is:

$$S = \{f(x)(x, -\lambda)| x = (u, v) \in \Omega\}$$

$$n(S(x)) = (\lambda \nabla f(x), f(x) + x \cdot \nabla f(x))$$
and the corresponding reflectance equation

\[ I(x) = \frac{\lambda I \cdot \nabla f(x) + c(x \cdot \nabla f(x) + f(x))}{\sqrt{\lambda^2 |\nabla f(x)|^2 + (x \cdot \nabla f(x) + f(x))^2}} \]  (7)

or rewritten in non-homogeneous form \((w = \ln(z))\)

\[ I(x) = \frac{\lambda I \cdot \nabla w(x) + c(x \cdot \nabla w(x) + 1)}{\sqrt{\lambda^2 |\nabla w(x)|^2 + (x \cdot \nabla w(x) + 1)^2}} \]  (8)

The SFS reflectance equations 6,8 are Hamilton-Jacobi type equations and in general they do not have smooth solutions [16] and require boundary conditions. There are two approaches to compute solutions for the SFS equations. One is to consider only the smooth solutions like in the work of Dupuis and Oliensis [31] or Kozera [76] but in practice there are no such solutions because of image noise, incorrect modeling (interreflections, non-Lambertian reflection,...) or errors in calibrations parameters (light,camera). An alternative approach is to compute viscosity solutions [89] (solution that is smooth almost everywhere or maximal weak solutions). The advantages of the theory of viscosity solutions are that it insures the existence of weak solutions if the image is continuous, it can characterize all solutions, and it guarantees that any solution is effectively computable. But they still require knowledge at image boundary, not a practical assumption.

Prados et al. extended the existing work of solving SFS based on viscosity solutions by introducing two algorithms for computing viscosity solutions in the orthographic case [37] and perspective [111] case based on the Legendre transform. They prove convergence even in the case of cast shadows when the image is discontinuous. In a later paper [112] they introduced a new class of weak solutions, called “singular discontinuous viscosity solutions” (SDVS), that guarantees the existence of a solution in a large class of situations and do not necessary require boundary conditions. Recently [110] they prove that for a finite point light source, when the attenuation factor is considered the Shape From Shading problem can be completely well-posed.

Most of the work in SFS deals with uniform Lambertian surfaces and given light source direction. Some extensions to non-Lambertian surfaces have been made. The work of Oren and Nayar [100] solves the case of rough diffuse surfaces such as sand and plaster using a statistical model of the surface geometry. Samaras and Metaxas [118] use a model based approach where SFS is incorporated as nonlinear holonomic constraints within a deformable model framework. They demonstrate that any type of illumination constraint can be incorporated in the proposed...
framework (if it is a differentiable function) thus extending the usual Lambertian assumption to modeling rough diffuse surfaces. The approach is extended in [119] where both the shape and light direction are alternatively estimated. The method can be used within orthographic or perspective assumptions. Some other methods that iteratively estimate illumination direction and shape are presented by Brooks and Horn [15] or Zheng and Chellappa [167]. In the later, the elevation of the illumination and surface albedo are estimated from image statistics considering the effect of self shadowing.

One other important extension that accounts for global illumination and interreflections is presented by Langer and Zucker [81]. In the classical SFS model, surface luminance depends mainly on the surface normal. They showed that, under diffuse light conditions ("sky on a cloudy day"), luminance depends primely on the amount of sky visible from each surface element. They formulate this dependency with an approximate constraint between surface luminance and aperture and provide an efficient algorithm for recovering depth from images. In the classical SFS formulation, assuming single distant light source, the local geometric constraints lie along the surface. In their proposed revised formulation of diffuse light local geometric constraints are found in a visibility field defined in the free space above the surface.

IV. PHOTOMETRIC STEREO

Photometric stereo recovers surface orientation (per-pixel based similar to SFS) from several images taken from the same viewpoint (under orthographic projection) with different but known illumination conditions. The classic formulation introduced by Woodham in the 80’s [153] recovers per-pixel surface orientation for a Lambertian object viewed under at least three different light conditions.

Denoting by $l_i \in \mathbb{R}^3, \ i = 1 \cdots n, \ n \geq 3$ the light directions and the corresponding image measurements for a given point by $I_i$ the photometric equations can be written as:

$$
\begin{bmatrix}
I_{1}^T \\
I_{2}^T \\
\vdots \\
I_{n}^T
\end{bmatrix}
\alpha \mathbf{n} =
\begin{bmatrix}
I_{1} \\
I_{2} \\
\vdots \\
I_{n}
\end{bmatrix}
$$

(9)

The normal $\mathbf{n} = (n_x, n_y, n_z)^T$ scaled with the albedo $\alpha$ can be easily calculated if the light directions span a 3D subspace. The albedo is then obtained by normalization. An equation system
of the above form can be written for each pixel, so the method independently reconstructs a normal and albedo for each pixel being. Therefore it can reconstruct a varying diffuse albedo as opposed to SFS method that assumed uniform albedo.

A. Uncalibrated photometric stereo

An interesting problem in photometric stereo is the uncalibrated case when both the light positions and the shape of a Lambertian object are unknown. The first to study the problem in the orthographic case was Belhumeur et al [5]. They proved that there is a 3 parameter family ambiguity (“generalized bas-relief” GBR ambiguity) in the reconstructed shape of a Lambertian object given a set of images taken under different unknown illumination conditions. The name comes from the flat sculptures called “bas-relief” popular in the antiquity that, when viewed from particular positions, are indistinguishable from full reliefs. Belhumeur et al demonstrate that: (1) the set of images under all illumination conditions including shadowing (illumination cones [50]) for a surface and the transformed ones are identical; (2) cast and attached shadows do not provide further information; (3) the set of motion fields produced by a surface and the transformed one are identical, so small motions do not provide any additional constraints.

The GBR ambiguity is formally introduced next. The object surface is parametrized as per-pixel depth map \( S(u, v) = (u, v, f(u, v))^T = (x, f(x))^T \). The reflectance equation for a surface point \( x \) with albedo \( a(x) \) and normal \( n = (-\nabla f, 1)/\sqrt{\nabla f \cdot \nabla f + 1} \) illuminated with a light source \( l \) is then (ignoring inter-reflections):

\[
I(x) = \Psi_{f,l}(x) a(x) n(x)^T l = \Psi_{f,l}(x) b(x) l
\]

where \( \Psi_{f,l} \) is a binary function that represents shadows (0 if \( x \) is in shadow and 1 otherwise). If the light source strength and direction are unknown \( b(x) \) can only be determined up to a \( 3 \times 3 \) linear transformation since the following equality holds for any invertible transformation \( A \):

\[
b l = (b A^{-1})(A l)
\]

so \( b A^{-1} \) and \( A l \) will also satisfy equation 10.

When imposing integrability constraints on the reconstructed surface \( f_{uv} = f_{vu} \) the linear
ambiguity is reduced to a 3 parameter family ambiguity GBR [5]:

\[
G = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
\mu & \nu & \lambda
\end{pmatrix}
\]  \hspace{1cm} (11)

\[
\tilde{f}(u, v) = \lambda f(u, v) + \mu u + \nu v
\]  \hspace{1cm} (12)

When \( \mu = \nu = 0 \) and \( \lambda > 0 \) the transformation is called “classical bas relief” transformation. A GBR transformation for a surface-albedo pair is defined by a transformation of the normal field and albedo:

\[
\mathbf{n} = \frac{G^{-T} \mathbf{n}}{|G^{-T} \mathbf{n}|}, \quad \bar{a} = a|G^{-T} \mathbf{n}|
\]  \hspace{1cm} (13)

while the corresponding light source transformation is:

\[
\bar{I} = GI
\]  \hspace{1cm} (14)

Assuming constant (or known) albedo or same light intensity for all the light sources the GBR ambiguity is restricted to a binary subgroup given by \( \lambda = \pm 1, \mu = \nu = 0 \) [5]. The in-out (convex-concave) ambiguity can be resolved using cast shadows.

The shape reconstruction methods (up to a linear ambiguity) are based on SVD that splits the images into components dependent on surface properties and light conditions [57], [41], [158]. Hayakawa [57] resolved the linear ambiguity by considering some special cases (e.g. set the rotation ambiguity in the SVD to identity). Yuille et al. [158] proved later that those constraints are quite restrictive. Fan and Wolff [41] use image ratios obtained from a stereo pair to eliminate the albedo (ignoring shadows) and estimate surface properties such as the sign of curvatures. The most general method presented by Yuille et al [158] generalizes the SVD approach to include ambient light and use knowledge about the class of objects to resolve the ambiguity. They also presented an iterative approach that, combined with the SVD, detects and eliminates shadows improving the quality of results.

The uncalibrated photometric stereo under perspective projection was studied by Kriegman and Bellhumeur [77]. They found a similar result to the one in the orthographic case. They showed that Lambertian surfaces that differ by 4 parameter family of projective transformations (GPBR) viewed under different illumination conditions (finite number of light sources) are shadow equivalent. Therefore they can only be reconstructed up to this 4D ambiguity. They also
present a method for reconstructing shape up to a GBR (in the case of orthographic camera) from attached shadows.

Simakov and Basri [4] generalized the uncalibrated photometric stereo for the case of general light, not restricted to finite number of light sources. Their work is based on the result that Lambertian objects under general light conditions can be represented using low order spherical harmonics [113], [3]. They recover the shape by fitting a low dimensional space that represents the actual images with a similar space that could be produced by the 3D scene. They also investigate to what extent a linear description of object images uniquely determine the surface normals and found that for the case of four harmonics it is a $4 \times 4$ linear transformation (Lorentz transform). When imposing integrability constraints on the surface they proved that their method is able to solve the GBR ambiguity as long as the first four harmonic images can be identified as dominant components of the image.

An interesting recent result was presented by Kriegman et al. [20] that integrated interreflections in the study. They showed that for general non-convex surfaces, interreflections completely resolve the GBR ambiguity but there is still a non-linear ambiguity for objects with translational symmetry.

B. Photometric stereo for non-Lambertian surfaces

The general assumption for Lambertian surfaces in photometric stereo can practically cause problems [128], [153] as real surfaces are seldom perfectly Lambertian. Therefore the algorithm will solve for a normal map which can lead to geometric distortions. Coleman and Jain [38] were the first to reconstruct non-Lambertian surfaces using photometric stereo. They assumed that the BRDF is a linear combination of a Lambertian and a specular component with limited angular support so specular components are treated as outliers and discarded. The traditional photometric stereo technique is than applied for the remaining diffuse pixels. Since then, different methods that account for a general BRDF in photometric stereo have been developed. They can be divided into approaches that separate specular pixels and treat them as outliers and approaches that explicitly represent the general BRDF by either a parametric or a non-parametric model. We next present the most representatives works in the literature.

A common approach in separating diffuse and specular components is to assumes dichromatic surfaces (surface whose reflectance can be represented using Shafer’s dichromatic model [127]).
Shafer’s model suggest that in the case of dielectrics (non-conductors) diffuse and specular components have different spectral distributions. Schluna and Witting [124] made the assumption that the surface is homogeneous dichromatic to separate the diffuse and specular parts using color histograms. A similar approach [122] avoids the restriction to homogeneous regions by using a large number of light source directions to compute a color histogram at each point. A recent interesting method applicable for dichromatic surfaces presented by Mallick et al [93] uses a data dependent rotation of RGB color space that allows easy isolation and removal of specular effects. These methods have the additional benefit of recovering the diffuse color at each point, but they are somewhat restrictive as they assume that the specular lobe is narrow relative to surface curvature.

The uncalibrated photometric stereo in the case of specular reflections was studied by Drbohlav and Sara [29]. They employ the consistent viewpoint constraint on a minimum of 4 normals of specular points to reduce the ambiguity. The method is a Lambertian photometric stereo where specularities are treated as outliers, so it will not be able, in general, to recover shape of non-Lambertian objects (with specular lobes) or recover their non-Lambertian reflectance. More general results are presented by Georghiades et al [51] that approximate specular reflectance with a Torrance-Sparrow parametric model. They proved that when considering specularities the ambiguity is resolved up to a binary convex/concave ambiguity) and results in a more accurate surface reconstruction. The GBR ambiguity is resolved by aligning the specular lobes in the acquired images with respect to light sources and cameras with the lobes derived from the TS model.

An innovative approach that recovers surfaces with general reflectance properties under unknown distant light (ignoring shadows and interrelations) is presented by Hertzmann and Seitz [58]. They use the observation that if a reference object with similar material and known geometry is introduced into the scene, the scene normals can be reconstructed using an orientation-consistency measurement. For surfaces with varying material properties, a full segmentation into different material types is also computed. In a later paper [145] they use the same consistency measurement for a volumetric reconstruction technique (see Section V-B). Recently [53] they eliminate the need of the reference object and instead estimate the BRDF for a small number of fundamental materials. This approach recovers not only the shape but also material BRDFs and weight maps in an alternative way, yielding compelling results for a wide variety of objects.
V. MULTI-VIEW METHODS

We present here reconstruction techniques that use multiple calibrated views of the same scene. Compared to the SFS and photometric stereo methods they are able to reconstruct full objects. Since they estimate 3D shape, not normals, discontinuities can be preserved. Multi-view methods are based on correspondence between images which, unlike in the case of photometric stereo, here it is not trivial. If feature correspondences are available the corresponding 3D feature positions can be calculated through triangulation (the well known stereo constraint). Therefore, if correspondences are available, the scene can be reconstructed without explicitly modeling the light or surface reflectance. Unfortunately, in practice, correspondence is a difficult ill-posed problem because of image noise, texture-less regions, depth discontinuities, occlusions, uncounted effects (s.a. specularities, transparency, interrelations). As a consequence some constraints have to be imposed in order to solve the reconstruction problem. Some of the most common ones are photo-consistency, local smoothness, generalized ordering, or uniqueness.

<table>
<thead>
<tr>
<th>Reflectance</th>
<th>Light</th>
<th>Cost function for surface point $x$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>stereo</strong></td>
<td>textured Lambertian</td>
<td>constant varying $SAD_{12}(x) = (I_1(P(N_5(x))) - I_2(P(N_5(x))))^2$, $N_5(x)$ neigh. of $x$</td>
</tr>
<tr>
<td></td>
<td>varying</td>
<td>$ZNCC_{12}(x) = N_1 N_2$, $N_i = \frac{I_i(P(N_5(x))) - I_i(P(N_5(x)))_{neigh.}}{I_i(P(N_5(x)))}$, $i = 1, 2$</td>
</tr>
<tr>
<td><strong>shading</strong></td>
<td>uniform</td>
<td>constant (SFS) $SFS_i(x) = I_i(P(x)) - E(x, n, \rho, l_i, v_i)$, $n$ normal, $\rho$ BRDF,</td>
</tr>
<tr>
<td></td>
<td>textured/uniform</td>
<td>varying (PS) $PS_i(x) = I_i(P(x)) - E(x, n, \rho, l_i, v_i)$, $l_i$ light, $v_i$ view dir.</td>
</tr>
</tbody>
</table>

**TABLE II**

**SUMMARY OF PHOTO-CONSISTENCY CONSTRAINTS FOR MULTI-VIEW RECONSTRUCTION**

The choice in modeling the reconstruction problem depends on the assumptions about surface material reflectance and light. Table II summarizes the most common cost functions (photo-consistency constraints) for stereo and shading ques. Most of the multi-view approaches assume that the scene is Lambertian and textured (has non-uniform albedo). In this case, correlation scores on small patches around feature points can be defined to solve the correspondence problem (**stereo constraints**). Some examples of correlation score include SSD (sum of squared differences), SAD (sum of absolute differences), and ZNCC (zero normalized cross correlation). For non-textured surfaces, **shading constraints** using explicit surface reflectance representation similar to the ones used for SFS or photometric stereo have to be defined in order to recover...
the shape. Like in the case of photometric stereo there also exist methods for reconstructing specular surfaces either by filtering out the specular pixels and applying the reconstruction for the remaining diffuse pixels or by explicitly modeling a general reflectance function \( E(x, n, l) \) (see Section V-F).

As mentioned in the introduction there are two distinct cases regarding illumination conditions: constant vs varying. The first one arises when the camera moves and the object is stationary with respect to light. In this case there is no image intensity variation for corresponding points on a Lambertian object (“multi-view shape from shading”). Therefore, for non-textured surfaces (ignoring cast shadows) the correlation score does not give any information about correspondences and the surface reflectance has to be explicitly modeled for recovering the shape (see [67] for a related discussion). Additional constraints compared to the one-view SFS problem are caused by occlusions or view-dependent specular effects. The second situation arises when the object moves (usually rotates) and the camera is stationary, resulting in different illumination conditions for different views (“multi-view photometric stereo”). In this case the reconstruction methods have to account for light variation. This can be done either by defining a robust correlation score (e.g. ZNCC [39], [40]) that works well for diffuse textured surfaces or by explicitly modeling surface reflectance with known or recovered light. In the case of photometric stereo type constraints (varying light) a varying albedo can be reconstructed while in the case of SFS type constraints (constant light), light cannot be deconvoluted from reflectance so the surface is usually assumed to have uniform (unit) albedo.

The surface representation as a disparity map is one of the first approaches to the multi-view reconstruction and is appealing because of its simplicity but it also suffers from several problems. Disparity is limited to image resolution and can produce aliased results so sub-pixel approaches have to be considered to increase the precision. Also potential smoothness terms (regularizers) are defined on image disparity and hence they are viewpoint dependent. Alternative methods overcome this problem by using a geometric object centered representation (e.g. volumetric grid, 3D mesh, level sets). In the following subsections we present the most common multi-view reconstruction techniques grouped based on surface representation (Subsections V-A, V-B, V-C). We will mainly focus on issues related to the representation for each case. We next discuss two popular choices for solving the optimization problem, one discrete technique (graph-cut in Subsection V-D) and variational techniques in Subsection V-E). The last part of this section
presents recent results and methods that reconstruct non-Lambertian surfaces.

A. Multi-view stereo

Computing a depth/disparity map for the pixels in one or several reference views is one of the most common approaches to the multi-view stereo problem. There is an ample literature in the area and a comprehensive survey and evaluation of two-frame dense stereo algorithms is presented by Scharstein and Szeliski [123]. Two view approaches deal with rectified images that have the epipolar lines aligned with the image scan lines making the search process easier (see [107], [48] for details on rectification). For two rectified images corresponding points are sitting on aligned scanlines (have the same vertical coordinate $v$) and the difference in the horizontal coordinate $u_l - u_r$, called disparity, is inverse proportional to the corresponding pixel depth. Therefore recovering disparity solves the reconstruction problem as depth is easily calculated from it.

Scharstein and Szeliski [123] classified the existing approaches into local and global methods. The local methods compute the corresponding point for a pixel in one image as the best candidate from the same scan line in the other image. The measure is a correlation score computed on a window centered around the pixels assuming local surface smoothness. As in practice it is difficult to choose the optimal window size and different approaches have been studied – adaptive windows, shiftable windows, compact windows. Through the correlation window the local methods assume certain smoothness but it is not globally imposed on the depth/disparity and neighboring pixels can have different disparities.

In contrast, the global methods use an explicit smoothness assumption and pose the stereo problem as a minimizing a global energy function that takes into account both matching and smoothness costs. Stereo matching scores are similar to the ones used for local methods (SSD, SAD, ZNCC) but in this case they are integrated over the whole surface. The smoothness cost or regularizer imposes that two neighboring pixels have similar depth (see the discussion about regularizers in Section V-E) and makes the reconstruction less sensitive to noise. The resulting discrete multi-view stereo optimization problem is NP hard but expressing it as a Markov Random Fields (MRF) energy optimization provides clean and computationally tractable formulation for which good approximate solutions exist (simulated annealing, mean-field annealing, graph cuts, iterative optimization, dynamic programming). The problem can also be formulated as continuous...
PDE evolution. Subsection V-D presents an overview of graph-cut techniques and Subsection V-E presents variational and level set formulations.

Stereo matching has been studied mostly in the context of small baseline and close to fronto-parallel planes. A small-baseline will degrade the accuracy of results and enforce the use of strong regularizers that introduce fronto parallel bias. Recently efficient methods for dealing with multiple large baseline images have been developed [138], [137], [49]. They produce accurate results and do not have the usual discretization and topological problems.

Most of the existing approaches produce a depth/disparity map with respect to a reference view. In general it is not the optimal solution as a single map does not cover the whole scene and parts of the available image information are not used. A better solution is to use multiple depth maps [141], [49] or a depth map relative to a low resolution base mesh [150].

Most stereo methods consider reconstructing only spatial disparities. Nevertheless, higher-order disparities such as orientation disparity are incredibly powerful for surfaces. The work of Li and Zucker [85], [86]. Traditional stereo implicitly use fronto parallel assumptions by imposing smooth disparity over a neighborhood. This assumption can cause problems when scene structure does not lie near the fronto-parallel plane when both orientation disparity as well as a positional disparity has to be considered. They propose a stereo reconstruction algorithm [85] takes advantage of both disparities by relating 2D differential structure (position, tangent, curvature) in the left and right images with the Frenet approximation of the 3D space curve. In a more recent paper [86] they exploit Cartan transport to combine geometric information from different surface normals in a neighborhood. By using both depth and normals integrated in the consistency score they showed that contextual information behaves like an extra geometric constraint for stereo correspondence.

Another group of related approaches approximate the surface with a collection of small patches (planar, quadratic) that are then integrated into a surface. An early method presented by Hoff and Ahuja [59] reconstructs patches (planar or quadratic) that are then interpolated to produce the final shape. Their patches do not cover the whole surface. The work is extended by Carceroni and Kutulakos [18] that propose a carving approach having as base element a surfel (dynamic surfel element). As this work is able to reconstruct surfaces with arbitrary BRDF it is presented in more detail later in Section V-F. A similar carving based on patches is presented by Zeng et al. [160]. They reconstruct each patch using graph cut optimization. Patches are carved out depending on
the resulting “quality” and the remaining ones are aggregated together using a distance field. In an earlier work [159] planar patches are build using a similar graph-cut optimization starting from a set of reliable feature points provided as input.

1) **Probabilistic stereo:** It is well known that images are noisy measurements. Assuming Gaussian noise the multi-view reconstruction problem can be posed in the Bayesian framework as finding the most likely surface that fits the given data (images). Beside offering a theoretical framework for modeling image noise, the Bayesian approach can also incorporate constraints like the smoothness (regularization) as priors and model outliers or visibility with hidden variables. Therefore they can account for occlusions and discontinuities. The process parameters are also learned during optimization as random variables that are either estimated or marginalized.

The Bayesian formulation has been applied to multi-view disparity based reconstruction [141], [137], [49], [140], [150]. The problem is formulated as a maximum a posteriori Markov Random Field (MAP-MRF) energy minimization. Following [137], let \( \mathcal{I} = I_1 \cdots I_n \) be the the input images (data) and \( f(x) \) the disparity value for \( x \) in the space of the reference image (\( I_1 \) for example). A binary hidden variable \( \mathcal{V}_{ix} \) models the visibility of reference pixel \( x \) in view \( i \). Together with disparity the corresponding ideal colors are also estimated in \( I^* \). Each image is modeled as a noisy measurement:

\[
I_i(P_i(x, f(x))) = I^*(x) + \epsilon
\]  

where \( P_i(x, f(x)) = P_i(x) \) is the function that maps the point \( x \) with estimated disparity \( f(x) \) in the space of image \( I_i \) and \( \epsilon \sim \mathcal{N}(0, \Sigma) \) models the image noise, \( \Sigma \) being the unknown covariance. Assuming calibrated images \( P_i \) can be easily expressed from the projection matrix. With the introduced notations, the reconstruction problem is posed as estimating the unknowns \( \theta = \{ f, I^*, \mathcal{V} \} \) given the data \( I_1 \cdots I_n \). Besides the introduced hidden variable \( \mathcal{V} \) has to be inferred during the reconstruction. In the Bayesian framework \( \theta \) is chosen as \( \theta^* = \arg \max_{\theta} p(\theta | \mathcal{I}) \) (it maximizes the posteriori probability). According to Bayes rule

\[
p(\theta | \mathcal{I}) = \frac{\int p(\mathcal{I} | \theta, \mathcal{V}) p(\theta | \mathcal{V}) p(\mathcal{V}) d\mathcal{V}}{p(\mathcal{I})}
\]  

\( L = p(\mathcal{I} | \theta, \mathcal{V}) \) represents the image likelihood and is modeled assuming images are independent measurements

\[
L = \prod_i \prod_x \mathcal{V}_{ix} \mathcal{N}(I_i(m_i(x)) | I^*(x), \Sigma)
\]
$p(\theta|\mathcal{V})$ is the prior on the model (disparity and scene colors). For the depth prior different expression have been utilized for example a data dependent anisotropic diffusion operator [137] gives good results. An interesting approach [49] that estimates multiple depth map uses a multiple depth prior that preserved depth discontinuities. An image-prior term was introduced by Fitzgibbon [44] enforcing images to look like natural images (using a database of patches extracted from real images).

$p(\mathcal{V})$ is the visibility prior usually uniform [137]. A geometric visibility prior that models occlusions is defined for multiple depth maps in [49].

$p(\mathcal{I})$ is the prior on the data (images) assumed constant and ignored.

Marginalizing $\mathcal{V}$ is computationally expensive but assuming that its probability function is peaked about a single value (Dirac function) leads to an Expectation-Maximization (EM) solution that iteratively estimates $\theta$ and $\mathcal{V}$. One of the problems with the EM algorithm is that in general it only finds a local minimum. A better approach is to use the Loopy Belief Propagation algorithm for approximating the posteriori probability [140], [150]. The LBP algorithm is based on local message passing techniques that solve inference problems on graphical models. One of its major advantages is that the message passing provides a time-varying adaptive smoothing mechanism that naturally deal with textureless regions and depth discontinuities. The disadvantage is that when the graph has cycles it is not guaranteed to converge.

**B. Volumetric reconstruction**

Volumetric reconstruction is an efficient way of solving the multi-view reconstruction problem. It was introduced in the 70’s in the field of medical imaging by Greenleaf [54] and used at the beginning for visual hull reconstruction problem. The approach consists of discretizing the space into a grid of voxels and then carving voxels that are not consistent within images. Assuming calibrated cameras, the image consistency is checked by re-projecting the voxel into the image space. For the visual hull reconstruction, a voxel is declared consistent if it projects inside all image silhouettes. The approach was later extended by replacing the binary consistency function with a color consistency (photo-consistency) measure, formally defined in [125], [78]. Thus the method reconstructs the volume that contains only the voxels that reproduces the image colors. The photo-consistency measure involves knowledge about surface reflectance and light conditions. Most of the approaches assume that the scene surface is Lambertian so consistency
can be checked based on color variance. In general there can be many 3D scenes consistent with a particular set of images so the reconstruction problem is ill-posed. The volume remaining after carving named the \textit{photo-hull}, represents the maximal shape consistent with the input images [78].

Voxel based algorithms use an object-based (geometric) representation, so they do not have the usual problems of image-based representations (e.g. viewpoint dependent regularization). Also compared to other surface based approaches (e.g. mesh, variational) they can handle complex structures by not making assumptions about scene planarity, continuity or topology. But the initial discretization limits the scene resolution and ignores regularization so the voxel based approaches tend to be more susceptible to noise. Also the visibility is not implicitly modeled, and is implied by the order of traversing the voxels that is therefore quite important. Another problem with current carving approach is that once a voxel is carved it cannot be recovered (holes get carved). As comprehensive reviews [35], [133] for volumetric methods exist in the literature, we give here just a short overview of existing approaches for completeness. Methods based on level sets are also variations of discrete 3D sampling, but, as they are derived from continuous formulations, they are discussed together with the variational methods in Section V-E.

The first proposed methods to solve the ordering problem was to use a plane that sweeps the scene from one side to the other [23]. This approach constrains the placement of the camera as lying on the same side of a plane. Seitz and Dyer [125] showed that it is possible to topologically sort the voxels for any camera configuration for which the scene volume lies outside the convex hull of the camera centers. The constraint on camera placement was eliminated using a multi-pass procedure [78], [26].

Some limitations of the early techniques was that the original photo-consistency score considered only color variations and ignored illumination variation or reflectance effects. Another possible cause of inaccuracies is discretization. Some efforts to improve the resolution considered adding a post-processing step that refines the model using optimization. Slabaugh et al [131] adopted an iterative post processing method that adds and removes surface voxels until the SSD between image and projected model is minimized using greedy simulated annealing. The refinement results in a reconstruction that is tighter fit to true scene geometry. Cross and Zisserman [24] use an SSD score computed on a quadratic local surface patch fitted in a neighborhood of surface voxels. Esteban and Schmitt [40], [39] adopt a voting scheme based on
ZNCC to find the potential depth of the surface at a particular 3D position. A triangular mesh is then used to extract the surface from the voxelized volume. The mesh is formulated like a 3D snake that moves toward the surface based on the gradient of the voxel volume, a smoothing term and the gradient of a silhouette based energy. The benefit of using a correlation score is that it can account for illumination variation for Lambertian parts and filter specular effects. An efficient voxel based optimization is presented by Vogiatzis et al [149]. In their work a photo-consistency cost functional is defined and discretized with a weighted graph. The optimization uses graph cuts that is known to converge in polynomial time (see Subsection ??). The algorithm uses the visual hull to infer occlusions and to constrain the topology of the scene.

1) Probabilistic framework: Another way to improve robustness is to adopt a probabilistic framework where a per-voxel probability is assigned based on appropriate likelihoods. Marginalizing all possible visibility configurations have to be considered but in practice the problem is approximated based on heuristics [14], [97], [9] or solved in a stochastic manner through hundreds of iterations [7]).

We next give a short formulation of probabilistic voxel carving [14]. Let $\mathcal{I} = I^i, i = 1, n$ be the input images and $\mathcal{P} = P^i$ the corresponding calibration matrices. The existence of a voxel at index $\{k, l, m\}$ is represented by a binary stochastic variable $\exists_{klm}$ that has value 1 when the voxel exist and 0 when the voxel is removed. Two additional variable define the voxel model – $\mathcal{V}_{klm}$ that describes the image data when voxel $klm$ exists and $\mathcal{W}_{klm}^i$ that describe image data for a missing voxel (one independent map per image). The probability of a voxel existing is then determined using Bayes’ rule (voxels are considered independent). The probability that a voxel exists is denoted $p(\exists = 1)$.

$$p(\exists_{klm} = 1|\mathcal{I}, \mathcal{P}) = \frac{P(\mathcal{I}, \mathcal{P}|\exists_{klm} = 1)p(\exists = 1)}{p(\mathcal{I}, \mathcal{P}|\exists_{klm} = 1)p(\exists = 1) + p(\mathcal{I}, \mathcal{P}|\exists_{klm} = 0)p(\exists = 0)}$$

(17)

where

$$p(\mathcal{I}, \mathcal{P}|\exists_{klm} = 0) = \prod_i \int_{\mathcal{W}_{klm}^i} p(I^i|\exists_{klm} = 0, \mathcal{W}_{klm}^i) p(\mathcal{W}_{klm}^i) d\mathcal{W}_{klm}^i$$

and

$$p(\mathcal{I}, \mathcal{P}|\exists_{klm} = 1) = \int_{\mathcal{V}_{klm}} \prod_i p(I^i|\exists_{klm} = 1, \mathcal{V}_{klm}) p(\mathcal{V}_{klm}) d\mathcal{V}_{klm}$$

We next present some extensions and improvements to the classic volumetric reconstruction algorithms. A detailed presentation of methods that reconstruct surfaces with general reflectance is presented in Section V-F.
2) **Non-opaque scenes:** The general assumption in voxel carving is that the scene has *opaque* voxels. Several extensions have been done that do not assume completely opaque scenes. Szeliski and Golland [142] designed a multi-step optimization procedure that assigns opacity to voxels (besides visibility and color). Partial opacity is assigned to mixed pixels - pixels that are at the occlusion boundary and have foreground/background colors. In the minimization they consider a smoothness constraint for color and opacities and a prior distribution for opacities. DeBonet and Viola [9] considered scenes with partially transparent voxels. They use an optimization method similar to [142] but more general as it allows arbitrary camera placement and reconstructs objects that are semitransparent. Based on the property that pixels are linear combinations of voxel colors along the rays, they recover the coefficients in the linear combination (responsibilities). The responsibilities are then used for determining opacities.

3) **Partial calibration:** Most approaches require that the input images are calibrated meaning that the internal parameters and relative position are known. Under this assumption the voxel reprojection in each image is known. Extensions have been made for intrinsically calibrated cameras [36] or weakly calibrated cameras [73], [117] that define a non-uniform voxel space based on the epipolar geometry.

4) **Large-scale environments:** Capturing large scale environments is difficult because of the different level of detail required to model different parts of the scene. An interesting approach to deal with this problem, called volumetric warping, is presented by Slabaugh et al [134]. The technique has spatially adaptive voxel size that increases with distance from camera and allows infinitely large scenes.

**C. Mesh-based reconstruction**

A 3D triangular mesh is one of the most common representations in computer graphics and some works have adopted it for surface reconstruction. Some obvious advantages of this representation are the capability of easily representing visibility, implicit normals, the possibility of fusing multiple types of information (e.g. shading, texture) and the use of existing graphic programs for efficient rendering. But methods have to explicitly handle self intersection and topological changes. They are usually avoided using a regularizer (smoothing term or internal energy). Many methods extract a mesh from the reconstructed shape as a last stage in the reconstruction but there exist also purely mesh-based methods. They can be seen as extensions...
in 3D of snake approaches from image segmentation and typically start with an initial mesh (usually the visual hull) and refine it based on the gradient of an error function with the goal to reproduce the input images. Like the variational approaches, a choice is to compute the error measure on the mesh (or texture space) or in the image space. The first case is easier to implement, but the algorithm has to explicitly account for image resampling. One other aspect to consider is whether the mesh has to be regular or not. An ideal mesh would adapt the size of triangles depending on the scene detail and discontinuities but in practice it is easier to implement a regular mesh. An example of a mesh that preserves discontinuities on an anisotropic mesh through a non-quadratic regularizer is presented by Lengagne and Fua [83].

The simplest photo-consistency measure for a triangle is, like in the case of other multi-view approaches, the color variance of the reprojected region in the input images. This corresponds to the case when the scene has Lambertian reflectance and there is no variation in illumination. Zhang and Seitz [163] use this measure with 3D points sampled on the triangular mesh. Vertices are moved based on the cumulated error for neighboring triangles. When the error is too big it is considered that the point is far from the surface and it is simply moved with a constant in the direction of the negative normal (toward the object). Mesh regularity is insured using Garland-Heckbert simplification algorithm that also reduces the resolution of the mesh in regions with little geometric detail.

A similar measure embedded into a variational formulation is used by Duan et al [30]. The error is computed on a tangent plane fitted around each vertex. The surface moves based on the solution of the PDE that involves a gradient term and a smoothing term dependent on surface curvature. An additional term similar to the one used by Zhang and Seitz moves the surface inward when there is little information about gradient or curvature. The mesh is initialized at low resolution such that it encloses the object and then further subdivided during the optimization. Unlike other mesh based approaches their implementation can handle topological changes. Another variational mesh-based reconstruction that can handle objects with general reflectance is presented by Birkbeck et al. [8]. They use a multi-resolution approach with numerical computed gradient (with finite differences).

Ishidoro and Sclaroff [65] compute the same color variance error but in texture space. For points with large error a search along epipolar lines is performed for better candidates. The recovered offsets are used to compute a free form deformation applied to the entire mesh. Slabaugh and Unal [132] define a 3D deformable surface called active polyhedron whose
vertices deform to minimize a regional boundary-based energy functional. The vertex motion is computed by integrating speed terms over polygonal faces of the surface. The resulting equations allow larger time steps compared to continuous active surfaces implemented with level sets.

Vogiatzis et al [148] adopt a probabilistic formulation where the error is computed for every image based on the all the other images warped in the space of this image. The optimization is performed using simulated annealing with different types of vertex mutations: random, gradient based, correlation based. Some other mutations like swapping/removing edges or introducing new edges are connected to the smoothness term. In a later work [150] they use a coarse mesh as base for sampling depth that is optimized using Loopy Belief Propagation.

As previously mentioned, a better measure for photo-consistency is a ZNCC score that implicitly accounts for light variation assuming Lambertian surfaces. Esteban and Schmitt [40], [39] propose a method that is more direct than the iterative ones by combining a carving approach with the mesh refinement (mentioned also with the volumetric approaches). A preliminary step computes correlation scores for a discretized volume that is converted to a gradient vector flow that will drive the mesh deformation together with a smoothing and a silhouette term.

Methods that reconstruct surfaces with general reflectance [47], [83], [156], [155] are discussed in Section V-F.

D. Graph cut techniques

Despite their binary nature, graph cuts have proven to be a useful multidimensional optimization tool that can efficiently solve the multi-view reconstruction expressed as a labeling problem [13], [63]. The discrete formulation of the multi-view stereo as labeling is the following: Assuming a surface discretization \( p = (p_u, p_v) \in P \) (e.g. pixels with disparities or voxels), find the set of labels \( f \) over a set \( L \) such that they minimize:

\[
E(f) = \sum_p \Phi_{data}(f_p) + \lambda \sum_{p,q \in \mathcal{N}} \Phi_{smooth}(f_p, f_q)
\]

where \( \mathcal{N} \) is set of pair of adjacent points on the surface.

We next review the definition of a graph cut and the connection between the minimum cut and the maximum flow problem. Then we present the main characteristics of graph cut methods in the context of surface reconstruction. There are two major advantages of using graph cuts for multi-view reconstruction. Firstly, they provide a framework for approximating continuous
hypersurfaces on N-dimensional manifolds. This geometric interpretation is used in computing optimal separating hypersurfaces [103], [149]. A second advantage is that they are powerful energy minimization tools for a large class of binary and non-binary energies [75], [13].

The graph cut problem can be formally formulated as follows (refer to Figure 2): Let $G = (\mathcal{V}, \mathcal{E})$ be an oriented graph which consists of a set of nodes $\mathcal{V}$ and a set of oriented edges $\mathcal{E}$. $\mathcal{V}$ contains two special terminal nodes a source $s$ and a sink $t$. Each edge $(p, q)$ has assigned a non-negative cost $w(p, q)$. A $s/t$ cut $C = \{S, T\}$ is defined as a partition of nodes into two disjoint sets $S$ and $T$ such that the source $s$ belongs to $S$ and the sink $t$ belongs to $T$ (see Figure 2 right). The cost of the cut is defined as the sum of boundary edges $(p, q)$ s.a. $p \in S$ and $q \in T$. The minimum cut problem is to find the cut that has the minimum cost among all the cuts.

An important result in combinatorial optimization is that the minimum cut can be solved by finding a maximum flow from source to sink. The maximum flow can be informally defined as the maximum amount of liquid that can be send from source to destination interpreting edges as pipes with capacity given by their weights. There exist algorithms to solve this problem in polynomial time (e.g. push-rebel algorithms [52] or augmented paths [45].

The geometric interpretation of graph-cuts as N-dimensional hypersurfaces makes them an attractive framework for the multi-view stereo problem. When the graph is fully embedded in the working geometric space, nodes, edges and their cost on the surface are defined such as the cost of a cut corresponds to the function that has to be optimized (surface functional).
One of the first applications of graph-cuts for hypersurface extraction in computer vision was stereo. Roy and Cox [116] and Ishikawa and Geiger [64] proposed similar methods where disparity maps are interpreted as separating hypersurfaces on 3D manifolds. Linear [116], [12], [17] or convex [64], [159] smoothing is insured by an anisotropic cut metric. For prohibiting folds in the cut, that would lead to unfeasible disparity maps, hard constraints are introduced as penalty terms [62], [12]. Recent work by Paris et al [159], [103] that designed a globally optimal graph cut technique with a convex smoothing term for an object centered surface representation. The graph cut algorithm extracts a distance field for small patches that are then integrated into a surface. An interesting geometry graph that incorporates exact silhouette constraints was recently presented by Sinha and Pollefeys [130]. Each valid cut separating source from sink yields a surface that exactly satisfies the silhouette constraints. A minimum cut will thus give the optimal surface when assigning costs for photo-consistency and shape priors to edges.

An important result presented by Boykov and Kolmogorov [11] establishes the connection between the cuts and hypersurfaces in continuous spaces. They how to build a grid graph and set the weights such that the cost of the cut is arbitrarily closed to the area of corresponding surface for any anisotropic Riemmanian metric. Their result showed that graph cuts can be used for extracting minimal surfaces (geodesics). The method was recently applied by Vogiatzis et al. [149] for a volumetric space. They use the visual hull of the object for inferring occlusions and constrain scene topology.

Graph cuts are also powerful multi-dimensional minimization tools. When the interaction energy is linear or convex the minimization problem can be exactly solved using a geometric graph [12], [64], [159]. But such energies do not preserve discontinuities. Kolmogorov and Zabih [75] provided a classification of binary energies that can be minimized with graph cuts but for the general multi-labeling problem there is no such result. One way to model discontinuities is to use the piecewise planar Potts model [109]: $\Phi_{smooth} = \lambda_{p,q}\rho(f_p \neq f_q)$, where $\rho$ is 1 if the argument is true and 0 otherwise. Unfortunately, energy minimization with Potts is NP hard but, as showed by Boykov et al. [13] there exist approximate solutions within factor 2 of the optimum. They also show that the algorithms proposed in [13] provide good approximations for a larger class of discontinuity preserving energies (metrics, regular interactions).
E. Variational and level set techniques

The variational formulation of surface reconstruction was first developed as a natural extension of curve and surface evolution methods known as snakes that were reformulated in the context of PDE driven evolving curves [71], [19] (geodesic active contours) in a form that could be applied for the stereo reconstruction problem [115], [43]. The multi-view stereo is formulated as the reconstruction of a surface that minimizes a stereo matching cost integrated over the surface. This results in a set of PDE’s which describe the motion of a surface in time. The desired surface is reconstructed by evolving an initial surface through the motion described by the PDE.

We first review some basic concepts in curve and surface evolution (following [66]) after which we present the variational and level set formulation for the multi-stereo problem. Some more recent extensions of the original formulation are discussed at the end.

1) Curve evolution: A regular curve is a differentiable map $C : [a, b] \rightarrow \mathbb{R}^2; C(p) = [x(p), y(p)]^T$, for which $C'(p) \neq 0, \forall p \in [a, b]$. We have the following definitions:

- **tangent**: $t = \frac{C'(p)}{|C'(p)|} = \frac{[x_p, y_p]^T}{\sqrt{x_p^2 + y_p^2}}$
- **normal**: $n = \frac{[-y_p, x_p]^T}{\sqrt{x_p^2 + y_p^2}}$
- **length of segment**: $s(p) = \int_a^p |C'_q|dq$
- **curvature**: $k = < C''_p, n >, \ s - \text{arclength param}$

A curve evolution is a family of curves indexed in time $t$ denoted $C(., t)$ that moves according to the following fwb (PDE):

$$C_t = < V, n > + < V, t > t$$

with the initial boundary condition:

$$C(., 0) = C_0(.)$$

where $V : \mathbb{R}^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}^2$ is a vector field referred as the speed function. The term $< V, t > t$ moves points along the tangent direction so they remain in the original curve. Therefore the fwb can be simplified to

$$C_t = < V, n >$$

In an optimization problem the choice of the speed function depends on the choice of the cost functional $\Phi$ that is being minimized. Consider $\Phi : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^+$. The energy of the curve is
then:
\[ E = \int_C \Phi(\mathbf{x}, \mathbf{n})ds \]  (20)

where \( s \) is an arclength parametrization for \( C \). The energy in equation 20 is intrinsic as \( ds \) makes it independent of the choice of parametrization. The gradient flow of the above optimization problem is given by the Euler Lagrange equations of equation 20. It can be shown [66], [42] that the minimization flow has the form:
\[ C_t = (k\Phi - \Phi_x, \mathbf{n}) - k < \Phi_n, \mathbf{n} > - \mathbf{t}^T \Phi_{nx} \mathbf{t} + k\mathbf{t}^T \Phi_{nn} \mathbf{t} \]  (21)

When the curve energy is not dependent on the normal \( E = \int_C \Phi(\mathbf{x})ds \) the flow is simplified to:
\[ C_t = (k\Phi - \Phi_x, \mathbf{n}) \]  (22)

2) Surface evolution: A similar theory can be derived for the case of regular surfaces. \( S \subset \mathbb{R}^3 \) is a regular surface if for each point \( P \in S \) there exist a neighborhood \( V \) and a map \( \mathbf{X} : U \rightarrow V \cup S \) of an open set \( U \subset \mathbb{R}^2 \) such that:
   1) \( \mathbf{X} \) is differentiable
   2) \( \mathbf{X} \) is homeomorphism (\( \exists \mathbf{X}^{-1} : V \cup S \rightarrow U \) continuous)
   3) \( \forall (u, v) \in U \) the differentiable \( d\mathbf{X}|_{(u,v)} : \mathbb{R}^2 \rightarrow \mathbb{R}^3 \) is one to one.

\( \mathbf{X} \) is called a parametrization of \( S \). Here are some basic definitions:

<table>
<thead>
<tr>
<th>tangent vector</th>
<th>( w \in \text{span}(\mathbf{X}_u, \mathbf{X}_v) = T_P(S) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>( \mathbf{n}_P = \frac{\mathbf{X}_u \times \mathbf{X}_v}{</td>
</tr>
<tr>
<td>area</td>
<td>( A(R) = \int_Q</td>
</tr>
<tr>
<td>curvature</td>
<td>( H = \frac{1}{2} \left( \frac{&lt;n, \mathbf{X}_{uu}&gt;}{&lt;\mathbf{X}_u, \mathbf{X}<em>v&gt;} + \frac{&lt;n, \mathbf{X}</em>{uv}&gt;}{&lt;\mathbf{X}_u, \mathbf{X}_v&gt;} \right) (\mathbf{X}_u \perp \mathbf{X}_v) )</td>
</tr>
<tr>
<td>second fundamental form</td>
<td>( \Pi_P(w) = -&lt;d\mathbf{n}_P(w), w&gt;, \ \forall w \in T_P(S) )</td>
</tr>
</tbody>
</table>

A geometric interpretation of the surface mean curvature \( H \) is presented in Figure 3. \( H \) at a point on the surface is the average of the principle curvatures at the point. For a given point on the surface, consider a plane going through that point and containing the surface normal at that point. The plane intersects the surface creating a curve that lies on both the surface and the plane. For any such plane, the normal curvature is defined as the curvature of the plane
curve at the particular surface point. The principle curvatures at the point are the maximum and minimum of all possible normal curvatures [82].

The surface evolution that minimizes a functional

\[ E = \int_S \Phi(X, n) dA \]  

has the following form

\[ S_t = \left( 2H\Phi - \Phi_X, n \right) \cdot n \] 

(24)

where

\[ \Phi_{nn} = \sum_{i=1}^{k} \lambda_i p_i p_i^T, \quad \lambda_i \in \mathbb{R}, \quad p_i \in \mathbb{R}^3 \]

and the simplified form for the case \( E = \int_S \Phi(X) dA \):

\[ S_t = \left( 2H\Phi - \Phi_X, n \right) n \]  

(25)

3) Level set methods: Different strategies have been applied for numerically solving the surface evolution problem. One of the best approaches is to use level sets, originally proposed...
by Osher and Sethian [101]. The power of this representation comes from its natural way of handling topology and occlusions.

The original idea behind the level sets is to represent the surface as the zero level set of a higher dimensional functional $\Psi(X, t)$:

\[
\begin{cases}
\Psi(X) < 0 & \text{for } X \in \Omega \\
\Psi(X) > 0 & \text{for } X \notin \bar{\Omega} \\
\Psi(X) = 0 & \text{for } X \in \partial \Omega = S
\end{cases}
\]

where $\Omega \subset \mathbb{R}^3$ is the region bounded by the surface $S$, $\bar{\Omega}$ is the closer of $\Omega$ and $\Phi$ is called the level set function. In this parametrization the defined surface entities have the form:

- normal: $n = -\frac{\nabla \Psi}{|\nabla \Psi|}$
- curvature: $H = k = \frac{1}{2} \nabla \cdot \left( \frac{\nabla \Psi}{|\nabla \Psi|} \right)$
- second fundamental form: $\Pi_p(w) = \frac{1}{|\nabla \Psi|} w^T \nabla^2 \Phi w$

Remember the surface evolution equation $S_t = F \mathbf{n}$ where $F$ is the speed, with an initial condition $S(0) = S_0$. Writing it for the level set function $\Psi$ (s.t. $\Psi(S(t), t) = 0$) we derive the following evolution equation:

$$\Psi_t = -F < \nabla \Psi, \mathbf{n}> = F|\nabla \Psi|$$ (26)

To evolve a surface that minimizes a cost functional over the surface $E = \int_S \Phi(X, \mathbf{n})$, the expression of $F$ is given by the Euler Lagrange equations (similar to equation 24 or 25).

A level set representation is an implicit representation that results in an Eulerian formulation of the surface evolution. As mentioned before, compared to the Lagrangian formulation defined on an explicit surface representation, it has the advantages of naturally handling topological changes. An intuitive explanation of the concept would be that in the case of explicit representation there is an external, fixed reference frame, whereas in the case of implicit representation the frame of reference is attached to the object (moving in time). Efficient numerical schemes have been proposed for the optimization [126], [102]. Unfortunately the convergence of the evolution equation in the presence of noise is unknown.

4) Cost functions: We discuss here some cost functional adopted for level sets implementations. Faugeras and Kerivan [43] were the first to apply a level set formulation for the multi-view stereo problem. The surface is represented as a disparity map with respect to the first image. They define three types of matching functional first based on SSD between corresponding
pixel intensities, the second and the third ones using a cross-correlation on a tessellation of the tangent plane parallel with the retinal plane or arbitrary (respectively). As an example, consider two images the surface represented as a disparity map with respect to the first image $I_1: S = (x, f(x)), x \in \Omega$. The SSD cost function for $n$ images has the following form:

$$\Phi(S) = \sum_{i=2}^{n} \int_{\Omega} (I_2(P(x, f(x))) - I_1(x))^2 dx$$  \hspace{1cm} (27)

Note that in this case the integration takes place over the image domain.

A similar SSD type cost functional is used by Robert and Deriche [115]. They design an regularizer that preserves depth discontinuities as a function of the gradient depth. In a later work [1] they extend the previous work to the weakly calibrated cameras and recover projective depth. To ensure convergence and avoid local minima the method is embedded into a linear scale space.

To account for illumination variations and filtering specular highlights Jin et al [69] used the median score of all possible two image ZNCC scores computed on a surface patch. An interesting level set formulation that is decoupled from the nature of the image similarity measure (photo-consistency) is presented by Pons et al [108]. They integrate the dissimilarity measures between each image and the predictive images coming from all other cameras. The resulting minimization flow is simpler than previous ones [43], [136]. The method is designed for dynamic scenes and reconstructs both 3D structure and scene flow. A matching cost for non-Lambertian surfaces is presented by Soatto et al [136]. It is based on the low rank constraint of a radiance tensor computed on the tangent plane to each pixel.

Another type of cost function considers light and shading information and is suitable for texture-less regions. Prados and Faugeras [112] provided a unified variational approach for the shape from shading problem and efficient numerical schemes to solve the optimization (discussed in Section III). A cost functional that incorporates both shading and texture information can be designed for objects that have both textured and non-textured regions [67], [47]. In a later work Jin et al designed a more stable SFS algorithm by introducing an additional vector field for representing normals. They also reconstruct light conditions together with shape. An example of cost function that accounts for shading is:

$$\Phi(S, \rho, L) = \sum_{i=1}^{n} \int_{S} (I_i(P(X)) - E(X, n, \rho, L, v_i))^2 dA$$  \hspace{1cm} (28)
where \( E(X, n, v, \rho, L) \) is the reflectance function defined like in equation 3 with \( \rho \) being the BRDF, \( L \) the light (direction and color), \( n_v \) the view direction and \( n \) the surface normal. Note that here the integration takes place over the surface (regularized the area).

Most of the level set approaches are used in the context of multi-view stereo as they compute a depth map with respect to a reference view. An example of a level set implementation for a voxel carving approach is presented in [135]. In this case the functional (variance based) is computed over image regions that observe a particular voxel.

There is one aspect that needs to be discussed relative to variational and level sets methods related to the space where the cost functional should be computed (image equation 27 vs surface equation 28). Integrating over the surface, like in the level set formulation has the advantage of making the problem intrinsic (independent on parametrization) and also provides automatic regularization (since the surface appears in the area form \( dA \) - multiplicative regularizer).

Researchers (e.g. Soatto et al. [136]) made the observation that the approach might suffer from over smoothness due to high order derivatives involved into unknowns and also could change the image variability depending on the surface. A better approach might be to integrate in image space and add a regularization term, more like in the snake formulation. A similar regularization term can be expressed as a prior in the probabilistic formulation for multi-view stereo (see Section V-A.1).

Different types of regularization terms have been proposed. A good review of existing diffusion and regularization terms for vector valued problems is presented by Weickert and Brox [152]. Most of the smoothness terms used for the surface evolution problem were studies first in the context of curve evolution. One approach is to uniformly smooth the solution but a good regularizer should smooth the homogeneous regions while conserving discontinuities. Authors have considered discontinuities in depth (on the recovered map) [115], [83] or, alternatively, discontinuities in image space [139], [83].

A quadratic regularization improves convergence but penalizes large variations of the function and isotropically smooths the solution. A non-quadratic regularizer over a bidimensional domain can be expressed as \( E_{\text{reg}} = \int \Phi(|\nabla z|)dxdy \). It was shown [115] that such a regularizer can be decomposed in two terms one in the direction of \( \nabla z \) and the other in the directional orthogonal to \( \nabla z \). Imposing the conditions to smooth homogeneous regions while preserving discontinuities leads to specific function \( \Phi \) (e.g. Rudin function \( \Phi(x) = x \), Tikhonov function \( \Phi(x) = x^2 \)).
A non-quadratic regularizer applied to a deformable mesh is also used in [83]. Gargallo and Sturm [49] designed an interesting multiple depth map prior for a probabilistic multi-view stereo reconstruction algorithm.

An alternative is to regularize the solution with respect to image discontinuities. Alvarez and Deriche [1] proposed a anisotropic linear operator based on the Nagel-Enkelmann diffusion operator. A similar regularization term is adopted by Strecha et al [138].

**F. Surfaces with general reflectance**

We present here methods that explicitly model surface reflectance under known or unknown light conditions. It is known that stereo fails to reconstruct untextured regions and additional ques like the ones used for shape from shading of photometric stereo have to be used for the reconstruction (e.g the shading scores in Table II or the cost function in equation 28). For this explicit light and reflectance models have been considered. Remember the reflectance equation that models the amount of light received by a camera at direction $V$, from a point $X$ with normal $n$ illuminated by a single infinite light $L = (l_1, l_2)$ (direction and color),

$$E(X, n, \rho, L, V) = \rho(X, n, \rho, l_1, v) l < ln >,$$

(29)

Like in the case of single view techniques (SFS and photometric stereo) non-Lambertian reflectance surfaces have been reconstructed from multiple views using a parametric or non-parametric model for the unknown BRDF $\rho(X, n, \rho, l_1, v)$. The simplest example is the Lambertian BRDF that can be modeled by a scalar. Unlike the single view approaches that reconstruct per-pixel normals that do not preserve surface discontinuities, multi-view approaches can directly reconstruct 3D structure thus preserving discontinuities. We start the presentation with natural extensions of SFS and photometric stereo for Lambertian scenes and then present some methods that compute or account for arbitrary reflectance. Table III presents a summary of some existing approaches in dealing with general reflectance based on the method used (filter, account, parametric model and non-parametric model). Light conditions are shown in column 2: variable if the light is changing with respect to the object (e.g. moving object), constant light is fixed with respect to the object, and the available knowledge about light conditions known, accounted, reconstructed. Column 3 presents the type of reflectance and whether the surface is assumed uniform or textured. For each method we also specified in the last column the surface
representation (if not mentioned a depth/disparity representation is implied) and the method used
in reconstruction (e.g. multi-view, stereo, PS or SFS type constraints).

For dealing with uniform Lambertian regions stereo information can be combined with SFS
or photometric information. One approach for merging SFS and binocular stereo was proposed
by Cryer et al [25]. The authors compute SFS and stereo individually and then merge them in
the frequency space. The merging uses low frequency components in the stereo and the high
frequency components from the shading implying that the errors from stereo would propagate
to the final model. So, in a sense the two methods do not complete each other as it would be
desired. Samaras and Metaxas [120] proposed an iterative multi-view shape from shading for
recovering a piece-wise uniform albedo and shape. Based on the stereo error, they automatically
segment the albedo map using a Minimum Description Length (MDL) based metric, to identify
areas suitable for SFS and to derive illumination information.

Another interesting approach that recovers light conditions together with shape and a non-
uniform albedo is presented by Zhang et al [161]. For initialization, sparse 3D points, normals
and camera positions are recovered using tracking. This information is then used for recovering
illumination conditions assuming that it can be modeled as a directional light and an ambient
term. The light conditions are changing from one view to the other (as the object is rotating
in front of a stationary camera and light) providing photometric stereo constraints to recover
per-pixel dense normals and albedo. Depth is integrated from the normals and aligned with the
previously recovered 3D feature points to generate the final surface. A similar idea of using
feature points for calibrating light position and albedo is proposed Vogiatzis et al. [147]. They
make use of frontier points that imply their normals to reconstruct varying light position and
albedo. The surface is then reconstructed based on this knowledge from shading constraints.

A similar approach that iteratively reconstructs shape, Lambertian reflectance and light was
recently presented by Lim et al. [87]. They also use a set of points tracked and reconstructed using
Tomasi-Kanade factorization but in addition to similar work [161], they use them to initialize a
piecewise linear surface. Uncalibrated photometric stereo is used to globally reconstruct normals
and light positions while the GBR ambiguity is solved by constraining the surface to be close
to the one reconstructed from tracked points. The paper also proves that the proposed algorithm
converges.

Jin et al [67] proposed a variational framework that uses the position information from a
<table>
<thead>
<tr>
<th>Method description</th>
<th>Light</th>
<th>Reflectance</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter specular pixels</td>
<td>var/known</td>
<td>fi iter spec</td>
<td>PS</td>
</tr>
<tr>
<td>Coleman and Jain’82 [38]</td>
<td>var/known</td>
<td>fi iter spec</td>
<td>PS</td>
</tr>
<tr>
<td>Barsky and Petrou’03 [2]</td>
<td>var/known</td>
<td>fi iter (dichr surf)</td>
<td>PS</td>
</tr>
<tr>
<td>Sato and Ikeuchi’94 [122]</td>
<td>var/known</td>
<td>fi iter (dichr surf)</td>
<td>PS</td>
</tr>
<tr>
<td>Mallick et al.’05 [93]</td>
<td>var/known</td>
<td>fi iter spec</td>
<td>uncalib PS</td>
</tr>
<tr>
<td>Drobhov and Sara’02 [29]</td>
<td>var/rec</td>
<td>fi iter (color hist)</td>
<td>multi/depth, stereo</td>
</tr>
<tr>
<td>Lin et al.’02 [88]</td>
<td>ct/acc</td>
<td>fi iter (2 images)</td>
<td>multi/depth, stereo</td>
</tr>
<tr>
<td>Samaras et al.’00 [120]</td>
<td>ct/known</td>
<td>account</td>
<td>multi/voxels, stereo+SFS</td>
</tr>
<tr>
<td>Account general BRDF</td>
<td>ct/acc</td>
<td>account</td>
<td>multi/voxels, stereo</td>
</tr>
<tr>
<td>Chhabra’01 [22]</td>
<td>ct/acc</td>
<td>account</td>
<td>multi/voxels, stereo</td>
</tr>
<tr>
<td>Magda et al.’01,’03 [92], [168]</td>
<td>var/acc</td>
<td>account</td>
<td>multi/depth, PS (Helmholtz)</td>
</tr>
<tr>
<td>Davis et al.’05 [27]</td>
<td>var/acc</td>
<td>account</td>
<td>multi/depth+PS/depth (light transport ct)</td>
</tr>
<tr>
<td>Parametric reflectance</td>
<td>ct/known</td>
<td>rough diffuse/unif</td>
<td>SFS</td>
</tr>
<tr>
<td>Oren and Nayar’95 [100]</td>
<td>ct/known</td>
<td>rough diffuse/unif</td>
<td>SFS</td>
</tr>
<tr>
<td>Samaras and Metaxas’98 [118]</td>
<td>ct/known</td>
<td>rough diffuse/unif</td>
<td>SFS</td>
</tr>
<tr>
<td>Samaras and Metaxas’03 [119]</td>
<td>ct/rec</td>
<td>rough diffuse/unif</td>
<td>SFS</td>
</tr>
<tr>
<td>Samaras et al. [120]</td>
<td>ct/rec</td>
<td>Lamb/tex, elim spec</td>
<td>stereo+SFS</td>
</tr>
<tr>
<td>Cryer et al. [25]</td>
<td>ct/known</td>
<td>Lamb/tex</td>
<td>stereo+SFS</td>
</tr>
<tr>
<td>Fua and Leclerc [47]</td>
<td>ct/known</td>
<td>Lamb/tex</td>
<td>multi/mesh, SFS+S+PS</td>
</tr>
<tr>
<td>Lengagne and Fua [83]</td>
<td>ct/known</td>
<td>Lamb/tex</td>
<td>multi/mesh, SFS+S+PS</td>
</tr>
<tr>
<td>Simakov et al. [124]</td>
<td>ct/rec(sph harm)</td>
<td>Lamb/unif</td>
<td>multi, PS</td>
</tr>
<tr>
<td>Jin et al. [68]</td>
<td>ct/rec</td>
<td>Lamb/unif</td>
<td>extension, autom bk substr</td>
</tr>
<tr>
<td>Weber et al. [151]</td>
<td>var/known</td>
<td>Lamb/tex</td>
<td>multi/voxels, PS, account shadows</td>
</tr>
<tr>
<td>Zhang et al. [161]</td>
<td>var/rec</td>
<td>Lamb/tex</td>
<td>multi, PS</td>
</tr>
<tr>
<td>Lim et al. [87]</td>
<td>var/rec</td>
<td>Lam/tex</td>
<td>multi, PS</td>
</tr>
<tr>
<td>Jin et al. [67]</td>
<td>var/known</td>
<td>Lamb/tex+unif</td>
<td>multi, SFS</td>
</tr>
<tr>
<td>Vogiatzis et al. [147]</td>
<td>var/rec</td>
<td>Lamb/unif</td>
<td>multi, PS + stereo</td>
</tr>
<tr>
<td>Non-parametric reflectance map</td>
<td>var/rec</td>
<td>TorrSpar</td>
<td>uncalib PS</td>
</tr>
<tr>
<td>Georghiades’03 [51]</td>
<td>var/known</td>
<td>Phong</td>
<td>stereo+PS</td>
</tr>
<tr>
<td>Lange’99 [80]</td>
<td>var/known</td>
<td>Phong/unif</td>
<td>multi/mesh, PS</td>
</tr>
<tr>
<td>Yu et al.’04 [156]</td>
<td>var/known</td>
<td>Phong</td>
<td>multi/mesh, PS</td>
</tr>
<tr>
<td>Carceneroni and Kutulakos’02 [18]</td>
<td>var/known</td>
<td>Phong</td>
<td>multi camera/surfels</td>
</tr>
<tr>
<td>Bonfort and Sturm’03 [10]</td>
<td>var/known</td>
<td>mirror</td>
<td>multi/voxels</td>
</tr>
<tr>
<td>Non-parametric reflectance map</td>
<td>var/acc</td>
<td>example sphere</td>
<td>multi PS</td>
</tr>
<tr>
<td>Seitz et al.’03,’04 [58], [145]</td>
<td>var/acc</td>
<td>rec. small no. fund mat</td>
<td>PS</td>
</tr>
<tr>
<td>Goldman et al.’05 [53]</td>
<td>var/acc</td>
<td>unif.</td>
<td>multi, PS</td>
</tr>
<tr>
<td>Lu and Little’99 [90]</td>
<td>var/known</td>
<td>refl tensor</td>
<td>multi/variats, SFS</td>
</tr>
<tr>
<td>Soatto et al.’03 [136]</td>
<td>ct/acc</td>
<td>VIRM /ct</td>
<td>multi/mesh, SFS</td>
</tr>
<tr>
<td>Yu et al.’04 [155]</td>
<td>ct/acc</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE III**
single light source to reconstruct objects that have been partitioned in two regions: textured and non-textured. They propose a cost functional that uses shading information in non-textured regions and use Faugeras and Keriven [43] functional for stereo in textured regions. In a later work [68] they also recover a model for illumination. They assume that the light can be modeled as a collection of point light sources, an ambient term, and a uniform hemispherical term. As the albedo cannot be separated from the light color the albedo is assumed to be white.

A more principled method in combining stereo and shading information is presented by Fua and Leclerc [47]. They assume known illumination and use shading information together with stereo for a mesh based reconstruction. The algorithm iteratively estimates the shape and albedo. The stereo and smoothness terms are weighted based on the amount of texture in each triangle such that the stereo term is dominant when there is texture and the shading term otherwise. They use conjugate gradient for optimization assuming good starting point. A more recent work [83] extends the previous approach by replacing regular meshes with anisotropic ones and quadratic regularization with non-quadratic one. To improve the robustness in the presence of corrupted data and occlusions they show how a priori shape differential constraints (min/max curvature, normals) can be included in the reconstruction.

Some other methods use only shading and light information for shape recovery. Part of the volumetric reconstruction techniques, the system presented by Weber et al [151] uses a turntable and three light sources but with calibrated positions. Their consistency measure for carving voxels is the residual after fitting normal and diffuse albedo to the observed intensities. During reconstruction they use the light position information to also account for shadows. A similar photo-consistency measure was used by Simakov et al [129] but without the restriction of having calibrated lights. Their method is an extension to multi-view of the previous work on uncalibrated photometric stereo [4]. They recover shape based on the property that convex Lambertian objects under arbitrary illumination conditions can be represented using low order spherical harmonics[113], [3].

For modeling surfaces with specular reflectance different approaches have been adopted. The simple one is to eliminate/separate specular pixels and apply the stereo reconstruction algorithm to the remaining diffuse pixels. When working with textured surfaces a cross correlation score (e.g. Table II) provides a good photo-consistency measure for Lambertian regions. This score can also be used to detect and eliminate specularities as outliers when using multiple images [39].
Bhat and Nayar [6] propose a three camera setup where stereo matching can be done between any three pairs of cameras. They select the pair used for stereo computation based on the quality of the correlation score.

An alternative approach is to try to separate diffuse and specular components based on color statistics. Lin and Shum [88] integrate color analysis and multi-baseline stereo in a single framework by using color histograms to characterize point color. Their measure can detect and eliminate specular reflections.

Chhabra [22] and Yang et al. [154] propose a robust correlation score that accounts for both diffuse and specular reflectance by analyzing structured variation on the observed colors. Under constant light, the color variation can be expressed as the sum of a diffuse and a specular term. Assuming white specular color the observations should be on a straight line in RGB from the diffuse color to the color of light. The correlation measure is therefore the distance from this line in RGB space. This approach will work for some shiny objects but not with mirror like surfaces. An alternative carving approach for completely specular objects is proposed by Bonfort and Sturm [10]. They use multiple views of reflection of a calibrated scene to reconstruct voxels of a general 3D perfectly specular surface (mirror).

Another interesting approach presented by Magda et al. [92] and extended by Zickler et al. [168] that resembles photometric stereo but uses a stereo pair of images and Helmholtz reciprocity to recover shape for objects with arbitrary BRDF. Helmholtz reciprocity states that interchanging the incoming and outgoing light directions should produce the same effect. Therefore the BRDF term is eliminated between any two pairs of images that have the light sources and camera positions interchanged and traditional multi-view stereo is performed with the results not being affected by specular highlights. The method was extended to a multi-view approach [168]. They also present a second method [92] that uses a constraint that radiance along a ray of light is constant and recovers depth from images that have a double coverage of the scene’s incident light field. The methods are quite powerful as they do not need any information about surface reflectance but the acquisition setup is somewhat constrained. A similar idea of using a photometric constraint that accounts for arbitrary BRDF and light conditions is used by Davis et al. [27]. They introduce an invariant called light transport constancy that can be used to formulate a rank constraint on multi-view stereo matching. The method require that multiple images are taken with different illumination conditions (varying intensity) but the same light sources.
Specular reflection can also be explicitly modeled either with a parametric or a non-parametric BRDF. For having an accurate estimation of a parametric BRDF the object has to be sampled under many different (known/recovered) light conditions. Lange [80] used Phong’s reflectance model [106] in a reconstruction method that merges binocular stereo and photometric stereo. The same reflectance model is used by Yu et al [156] in an iterative algorithm that alternates shape computation with reflectance estimation. They represent the shape using a 3D mesh with spherical parametrization, assuming star-shaped objects. Both algorithms assume that the illumination conditions are known. Carceroni and Kutulakos [18] design a carving algorithm that works with a dynamic surfel element that incorporates all the information about a small surface patch - shape, reflectance, instantaneous motion. They rely on Phong reflectance model and estimate the reflectance parameters of each surfel assuming varying albedo and a uniform specular component.

A general BRDF can also be represented as a map that contains samples of actual measurements. Due to the high dimensionality (4D if wavelength is ignored) it is difficult to build such a map [157], [98] and various heuristic have been adopted.

Lu and Little [90] estimate a non-parametric BRDF from images of a rotating object (with uniform material) illuminated under a collinear light source. They show that in this case the diffuse and specular reflectance are functions of only one variable, the incident angle between surface normal and light. An iterative refinement algorithm is initialized using a precomputed reflectance fitted to a set of singular points whose geometry is estimated first. Then the method alternates between surface refinement and reflectance estimation. Yu et al [155] designed a system that estimates the shape of surfaces with general reflectance viewed under unknown but constant light conditions. They reconstruct a view independent reflectance map (VIRM) that combines both reflectance and illumination. The map has two 2D components one for diffuse pixels and the other for specular pixels. The light and reflectance component cannot be separated from VIRM but it can be used to render other shapes under the same illumination.

If an object with known geometry and similar material (e.g. a sphere) is introduced in the scene it can be used as reference to lookup surface orientations with similar observed colors [58], [145]. A multi-view volumetric reconstruction approach where a voxel is carved if the colors do not correspond to the looked up orientation. The idea can be extended to multiple materials when a linear combination of different material colors at a particular orientation is fitted to the observed colors. Recently [53], they eliminate the need of the reference object and instead
estimate the BRDF for a small number of fundamental materials.

Soatto et al [136] represent the reflectance properties at a point by the radiance tensor that is computed by sampling image intensities on a tessellation of the tangent plane. They formulate the reconstruction in a variational framework using a rank constraint on the radiance tensor field for measuring photo-consistency. The algorithm alternatively estimates both shape and radiance.

VI. DISCUSSION

We have reviewed different methods for reconstructing shape from images. There are many approaches different mainly based on the cost function (photo-consistency measure) and the surface representation.

Shape from shading works with texture-less regions when light information is available. Photometric stereo recovers both albedo and shape assuming known light variation, while for SFS the albedo is assumed uniform (usually white) photometric stereo techniques can reconstruct varying albedo. Uncalibrated photometric stereo recovers light together with shape but only up to a GBR ambiguity [5]. Some extensions have been done to reduce the ambiguity [51], [20]. Stereo matching cost works well for textured regions. When objects have both textured and non-textured regions a combination of stereo and shading has to be used [47], [67], [161]. This technique has the additional advantage of reconstructing albedo together with shape.

Classical algorithms have been design for Lambertian surfaces. Different approaches have been adopted for dealing with objects with non-Lambertian reflectance. The simplest one is to filter out (separate) specular pixels and apply the algorithm for diffuse pixels. Other estimate a parametric or non-parametric BRDF together with the shape. Recovering a parametric BRDF requires views under different known illumination conditions. Non-parametric models can instead capture reflectance maps (assuming uniform albedo) under constant light conditions [155], [136] but without being possible to separate light and reflectance.

Some observations can be made regarding the use of light information. Ideally methods should work in unknown light. Light variation for textured Lambertian objects can be accounted by using robust correlation score. But, light information is essential in reconstructing non-textured regions where stereo correlation scores do not work. Most methods assume known light thus constraining the capture to controlled laboratory settings (with one or several point light sources). Other recover illumination conditions alternatively with shape and reflectance [5], [119], [161],
But, still the assumptions on light setting (e.g. unique or finite number of light sources) limit the generality and applicability in unstructured or outdoor settings. A more general light can be captured using spherical harmonic coefficients [129] or using a special physically sensing device (light probe). As mentioned before, when the light is constant with respect to the scene, it can be captured in the reflectance map thus eliminating the need for light calibration [155], [136]. But for recovering reflectance parameters some knowledge of illumination conditions is necessary. Even for recovering albedo some knowledge of light is necessary to undo the shadows. As exceptions we mentioned here two approaches that work under general unknown light variation. One reconstructs objects with general BRDF under general light conditions if the camera and light are in particular configurations with respect to each other [92], [168] (e.g. Helmholtz reciprocity). An interesting but rather unpractical method that accounts for light variation on non-textured surfaces with general reflectance uses a reference sphere with similar material that is introduced in the scene [58], [145].

Regarding different choices of surface representations we have discussed some advantages and disadvantages along with the review. A depth/disparity map is suitable for single view techniques (shape from shading or photometric stereo) or binocular stereo. In the case of multi-view reconstruction this representation becomes less appealing. A single depth map could be used but then the reconstruction can become biased by the chosen reference camera. An alternative is to use multiple depth maps with respect to each camera, but then merging and consistency could become complicated. In either approaches handling visibility is problematic.

Object centered representations (mesh, voxels, level sets) are more suitable for multiple view reconstruction. Voxel representations are simple to implement and can handle complex structures without making assumptions about surface topology or continuity. But they have no way of representing normals needed for recovering shading information. Nevertheless, methods have been developed to take advantage of light information while using a voxel representation [151], [145]. Meshes and level sets give readily computable normals and easy way to calculate visibility. Level sets naturally handles topology changes while mesh based representations might have difficulties in dealing with topological changes. Meshes and level sets can also easily integrate regularization or smoothing terms compared to voxel grids.

Some other remarks can be made regarding the optimization method used in global multi-view reconstruction problems. Most methods (level set, variational) will compute only a local
minimum thus depending on initialization. Global convergent methods have been designed for
graph cuts [103] but for a restrictive type of energies.

We have seen that much progress have been achieved in shape modeling in the last decade.
Most effort has been made for eliminating the restrictive assumptions required by early techniques
(e.g. uniform Lambertian surface assumed for SFS algorithms, or highly textured surfaces for
stereo methods). Also, researchers worked on improving the accuracy of reconstructed model and
some current techniques (e.g. [40]) are able to produce models comparable in accuracy to laser
scanned models. But, one of the major challenging for image-based reconstruction methods is
noise. There exist ways of coping with image noise, like designing robust correlation scores but
many of the algorithms can become unstable due to noise. For example there is no study of the
convergence of the level set methods in the presence of noise and their numerical stability has to
be carefully addressed [101]. One other challenging problem are the specular surfaces. Still most
of the methods assume Lambertian scenes, that limit their applicability as most surfaces in the
real world cannot be approximated with a Lambertian model. One good way to deal with this
is to filter specular pixels and good methods have been designed for dichromatic surfaces (e.g.
[93]). But many applications require that both shape and reflectance are reconstructed so the
object/scene can be correctly integrated into a virtual worlds and rendered in new illumination
conditions. For estimating reflectance information about illumination conditions is required. This
is a major limitation and most current system assume calibrated light and therefore they work
only in carefully designed lab setups. But, despite these limitations, we believe that soon image-
based modeling systems are getting to a stage when they can compete with traditional modeling
(laser scans).

REFERENCES

[2] S. Barsky and M. Petrou. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of
1999.


