Full text available at: http://dx.doi.org/10.1561/0600000007

3D Reconstruction from Multiple Images
Part 1: Principles

3D Reconstruction from Multiple Images Part 1: Principles

Theo Moons

Hogeschool — Universiteit Brussel
Brussel, B-1000, Belgium
Theo.Moons@hubrussel.be

Luc Van Gool

Katholieke Universiteit Leuven
Leuven, B-3001, Belgium
Luc.VanGool@esat.kuleuven.be

ETH Zurich
Zurich, CH-8092, Switzerland
vangool@vision.ee.ethz.ch

Maarten Vergauwen

GeoAutomation NV
Leuven, B-3000, Belgium
maarten.vergauwen@geoautomation.com

now
the essence of knowledge
Boston – Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is T. Moons, L. Van Gool and M. Vergauwen, 3D Reconstruction from Multiple Images Part 1: Principles, Foundations and Trends® in Computer Graphics and Vision, vol 4, no 4, pp 287–404, 2008

ISBN: 978-1-60198-284-1

Foundations and Trends® in Computer Graphics and Vision
Volume 4 Issue 4, 2008

Editorial Board

Editorial Scope

Foundations and Trends® in Computer Graphics and Vision will publish survey and tutorial articles in the following topics:

- Rendering: Lighting models; Forward rendering; Inverse rendering; Image-based rendering; Non-photorealistic rendering; Graphics hardware; Visibility computation
- Shape: Surface reconstruction; Range imaging; Geometric modelling; Parameterization;
- Mesh simplification
- Animation: Motion capture and processing; Physics-based modelling; Character animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape Representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and Video Retrieval
- Video analysis and event recognition
- Medical Image Analysis
- Robot Localization and Navigation

Information for Librarians
Foundations and Trends® in Computer Graphics and Vision, 2008, Volume 4, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Foundations and Trends® in
Computer Graphics and Vision
Vol. 4, No. 4 (2008) 287–404

DOI: 10.1561/0600000007

now
the essence of knowledge

3D Reconstruction from Multiple Images Part 1: Principles

Theo Moons[1], Luc Van Gool[2,3], and Maarten Vergauwen[4]

[1] *Hogeschool — Universiteit Brussel, Stormstraat 2, Brussel, B-1000, Belgium, Theo.Moons@hubrussel.be*
[2] *Katholieke Universiteit Leuven, ESAT — PSI, Kasteelpark Arenberg 10, Leuven, B-3001, Belgium, Luc.VanGool@esat.kuleuven.be*
[3] *ETH Zurich, BIWI, Sternwartstrasse 7, Zurich, CH-8092, Switzerland, vangool@vision.ee.ethz.ch*
[4] *GeoAutomation NV, Karel Van Lotharingenstraat 2, Leuven, B-3000, Belgium, maarten.vergauwen@geoautomation.com*

Abstract

This issue discusses methods to extract three-dimensional (3D) models from plain images. In particular, the 3D information is obtained from images for which the camera parameters are unknown. The principles underlying such uncalibrated structure-from-motion methods are outlined. First, a short review of 3D acquisition technologies puts such methods in a wider context and highlights their important advantages. Then, the actual theory behind this line of research is given. The authors have tried to keep the text maximally self-contained, therefore also avoiding to rely on an extensive knowledge of the projective concepts that usually appear in texts about self-calibration 3D methods. Rather, mathematical explanations that are more amenable to intuition are given. The explanation of the theory includes the stratification

of reconstructions obtained from image pairs as well as metric reconstruction on the basis of more than two images combined with some additional knowledge about the cameras used. Readers who want to obtain more practical information about how to implement such uncalibrated structure-from-motion pipelines may be interested in two more Foundations and Trends issues written by the same authors. Together with this issue they can be read as a single tutorial on the subject.

Contents

Preface

Welcome to this Foundations and Trends tutorial on three-dimensional (3D) reconstruction from multiple images. The focus is on the creation of 3D models from nothing but a set of images, taken from unknown camera positions and with unknown camera settings. In this issue, the underlying theory for such "self-calibrating" 3D reconstruction methods is discussed. Of course, the text cannot give a complete overview of all aspects that are relevant. That would mean dragging in lengthy discussions on feature extraction, feature matching, tracking, texture blending, dense correspondence search, etc. Nonetheless, we tried to keep at least the geometric aspects of the self-calibration reasonably self-contained and this is where the focus lies.

The issue consists of two main parts, organized in separate sections. Section 1 places the subject of self-calibrating 3D reconstruction from images in the wider context of 3D acquisition techniques. This section thus also gives a short overview of alternative 3D reconstruction techniques, as the uncalibrated structure-from-motion approach is not necessarily the most appropriate one for all applications. This helps to bring out the pros and cons of this particular approach.

Section 2 starts the actual discussion of the topic. With images as our key input for 3D reconstruction, this section first discusses how we can mathematically model the process of image formation by a camera, and which parameters are involved. Equipped with that camera model, it then discusses the process of self-calibration for multiple cameras from a theoretical perspective. It deals with the core issues of this tutorial: given images and incomplete knowledge about the cameras, what can we still retrieve in terms of 3D scene structure and how can we make up for the missing information. This section also describes cases in between fully calibrated and uncalibrated reconstruction. Breaking a bit with tradition, we have tried to describe the whole self-calibration process in intuitive, Euclidean terms. We have avoided the usual explanation via projective concepts, as we believe that entities like the dual of the projection of the absolute quadric are not very amenable to intuition.

Readers who are interested in implementation issues and a practical example of a self-calibrating 3D reconstruction pipeline may be interested in two complementary, upcoming issues by the same authors, which together with this issue can be read as a single tutorial.

1

Introduction to 3D Acquisition

This section discusses different methods for capturing or 'acquiring' the three-dimensional (3D) shape of surfaces and, in some cases, also the distance or 'range' of the object to the 3D acquisition device. The section aims at positioning the methods discussed in the sequel of the tutorial within this more global context. This will make clear that alternative methods may actually be better suited for some applications that need 3D. This said, the discussion will also show that the kind of approach described here is one of the more flexible and powerful ones.

1.1 A Taxonomy of Methods

A 3D acquisition taxonomy is given in Figure 1.1. A first distinction is between *active* and *passive methods*. With active techniques the light sources are specially controlled, as part of the strategy to arrive at the 3D information. Active lighting incorporates some form of temporal or spatial modulation of the illumination. With passive techniques, on the other hand, light is not controlled or only with respect to image quality. Typically passive techniques work with whichever reasonable, ambient light available. From a computational point of view, active methods

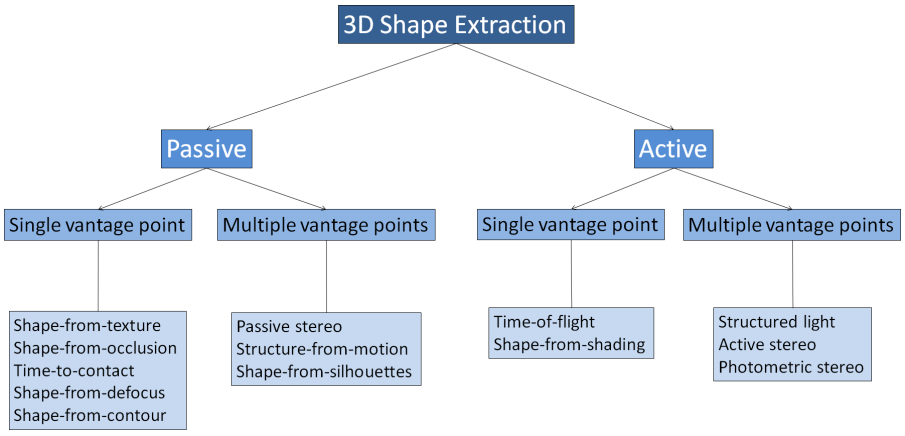

Fig. 1.1 Taxonomy of methods for the extraction of information on 3D shape.

tend to be less demanding, as the special illumination is used to simplify some of the steps in the 3D capturing process. Their applicability is restricted to environments where the special illumination techniques can be applied.

A second distinction is between the number of vantage points from where the scene is observed and/or illuminated. With *single-vantage methods* the system works from a single vantage point. In case there are multiple viewing or illumination components, these are positioned very close to each other, and ideally they would coincide. The latter can sometimes be realized virtually, through optical means like semi-transparent mirrors. With *multi-vantage systems*, several viewpoints and/or controlled illumination source positions are involved. For multi-vantage systems to work well, the different components often have to be positioned far enough from each other. One says that the 'baseline' between the components has to be wide enough. Single-vantage methods have as advantages that they can be made compact and that they do not suffer from the occlusion problems that occur when parts of the scene are not visible from all vantage points in multi-vantage systems.

The methods mentioned in the taxonomy will now be discussed in a bit more detail. In the remaining sections, we then continue with the more elaborate discussion of passive, multi-vantage structure-from-motion (SfM) techniques, the actual subject of this tutorial. As this

overview of 3D acquisition methods is not intended to be in-depth nor exhaustive, but just to provide a bit of context for our further image-based 3D reconstruction from uncalibrated images account, we do not include references in this part.

1.2 Passive Triangulation

Several multi-vantage approaches use the principle of *triangulation* for the extraction of depth information. This also is the key concept exploited by the self-calibrating structure-from-motion (SfM) methods described in this tutorial.

1.2.1 (Passive) Stereo

Suppose we have two images, taken at the same time and from different viewpoints. Such setting is referred to as *stereo*. The situation is illustrated in Figure 1.2. The principle behind stereo-based 3D reconstruction is simple: given the two projections of the same point in the world onto the two images, its 3D position is found as the intersection of the two projection rays. Repeating such process for several points

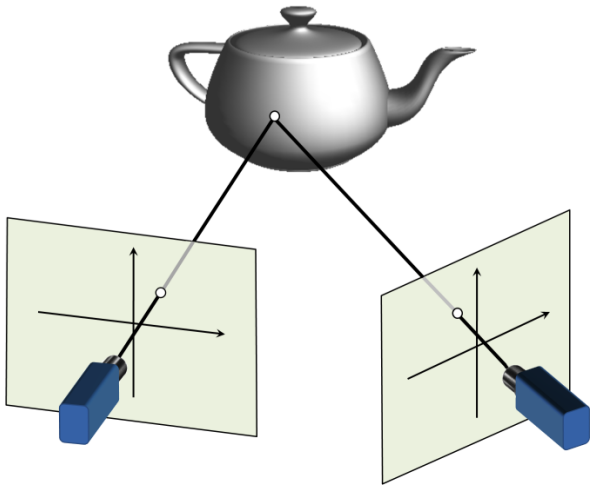

Fig. 1.2 The principle behind stereo-based 3D reconstruction is very simple: given two images of a point, the point's position in space is found as the intersection of the two projection rays. This procedure is referred to as 'triangulation'.

yields the 3D shape and configuration of the objects in the scene. Note that this construction — referred to as *triangulation* — requires the equations of the rays and, hence, complete knowledge of the cameras: their (relative) positions and orientations, but also their settings like the focal length. These camera parameters will be discussed in Section 2. The process to determine these parameters is called (*camera*) *calibration*.

Moreover, in order to perform this triangulation process, one needs ways of solving the correspondence problem, i.e., finding the point in the second image that corresponds to a specific point in the first image, or vice versa. Correspondence search actually is the hardest part of stereo, and one would typically have to solve it for many points. Often the correspondence problem is solved in two stages. First, correspondences are sought for those points for which this is easiest. Then, correspondences are sought for the remaining points. This will be explained in more detail in subsequent sections.

1.2.2 Structure-from-Motion

Passive stereo uses two cameras, usually synchronized. If the scene is static, the two images could also be taken by placing the same camera at the two positions, and taking the images in sequence. Clearly, once such strategy is considered, one may just as well take more than two images, while moving the camera. Such strategies are referred to as structure-from-motion or SfM for short. If images are taken over short time intervals, it will be easier to find correspondences, e.g., by tracking feature points over time. Moreover, having more camera views will yield object models that are more complete. Last but not least, if multiple views are available, the camera(s) need no longer be calibrated beforehand, and a self-calibration procedure may be employed instead. Self-calibration means that the internal and external camera parameters (cf. Section 2.2) are extracted from images of the unmodified scene itself, and not from images of dedicated calibration patterns. These properties render SfM a very attractive 3D acquisition strategy. A more detailed discussion is given in the following sections.

1.3 Active Triangulation

Finding corresponding points can be facilitated by replacing one of the cameras in a stereo setup by a projection device. Hence, we combine one illumination source with one camera. For instance, one can project a spot onto the object surface with a laser. The spot will be easily detectable in the image taken by the camera. If we know the position and orientation of both the laser ray and the camera projection ray, then the 3D surface point is again found as their intersection. The principle is illustrated in Figure 1.3 and is just another example of the triangulation principle.

The problem is that knowledge about the 3D coordinates of one point is hardly sufficient in most applications. Hence, in the case of the laser, it should be directed at different points on the surface and each time an image has to be taken. In this way, the 3D coordinates of these points are extracted, one point at a time. Such a 'scanning'

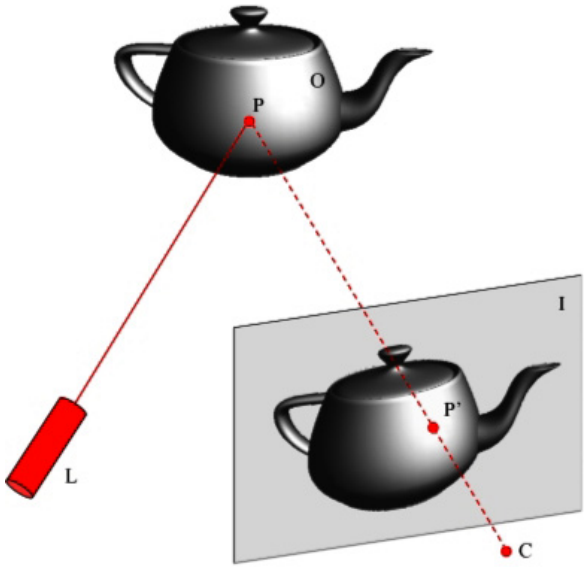

Fig. 1.3 The triangulation principle used already with stereo, can also be used in an active configuration. The laser L projects a ray of light onto the object O. The intersection point P with the object is viewed by a camera and forms the spot P' on its image plane I. This information suffices for the computation of the three-dimensional coordinates of P, assuming that the laser-camera configuration is known.

process requires precise mechanical apparatus (e.g., by steering rotating mirrors that reflect the laser light into controlled directions). If the equations of the laser rays are not known precisely, the resulting 3D coordinates will be imprecise as well. One would also not want the system to take a long time for scanning. Hence, one ends up with the conflicting requirements of guiding the laser spot precisely *and* fast. These challenging requirements have an adverse effect on the price. Moreover, the times needed to take one image per projected laser spot add up to seconds or even minutes of overall acquisition time. A way out is using special, super-fast imagers, but again at an additional cost.

In order to remedy this problem, substantial research has gone into replacing the laser spot by more complicated patterns. For instance, the laser ray can without much difficulty be extended to a plane, e.g., by putting a cylindrical lens in front of the laser. Rather than forming a single laser spot on the surface, the intersection of the plane with the surface will form a curve. The configuration is depicted in Figure 1.4. The 3D coordinates of each of the points along the intersection curve

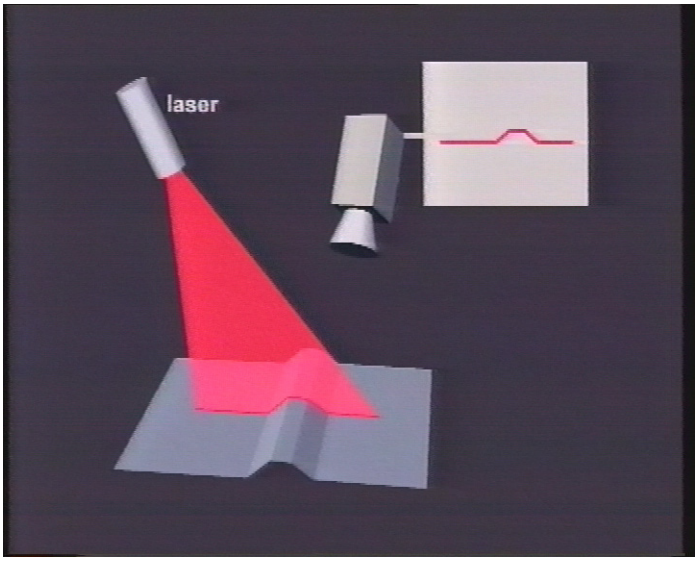

Fig. 1.4 If the active triangulation configuration is altered by turning the laser spot into a line (e.g., by the use of a cylindrical lens), then scanning can be restricted to a one-directional motion, transversal to the line.

can be determined again through triangulation, namely as the intersection of the plane with the viewing ray for that point. This still yields a unique point in space. From a single image, many 3D points can be extracted in this way. Moreover, the two-dimensional scanning motion as required with the laser spot can be replaced by a much simpler one-dimensional sweep over the surface with the laser plane.

It now stands to reason to try and eliminate any scanning altogether. Is it not possible to directly go for a dense distribution of points all over the surface? Unfortunately, extensions to the two-dimensional projection patterns that are required are less straightforward. For instance, when projecting multiple parallel lines of light simultaneously, a camera viewing ray will no longer have a single intersection with such a pencil of illumination planes. We would have to include some kind of code into the pattern to make a distinction between the different lines in the pattern and the corresponding projection planes. Note that counting lines has its limitations in the presence of depth discontinuities and image noise. There are different ways of including a code. An obvious one is to give the lines different colors, but interference by the surface colors may make it difficult to identify a large number of lines in this way. Alternatively, one can project several stripe patterns in sequence, giving up on using a single projection but still only using a few. Figure 1.5 gives a (non-optimal) example of binary patterns. The sequence of being bright or dark forms a unique binary code for each column in the projector. Although one could project different shades of gray, using binary (i.e., all-or-nothing black or white) type of codes

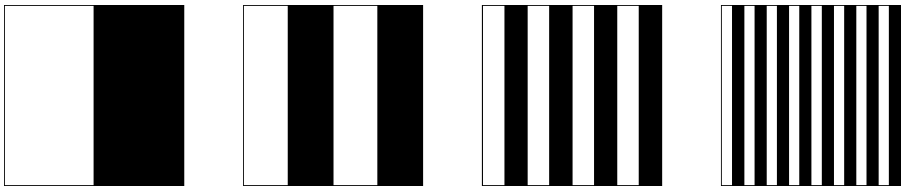

Fig. 1.5 A series of masks that can be projected for active stereo applications. Subsequent masks contain ever finer stripes. Each of the masks is projected and for a point in the scene the sequence of black/white values is recorded. The subsequent bits obtained that way characterize the horizontal position of the points, i.e., the plane of intersection (see text). The resolution that is required (related to the width of the thinnest stripes) imposes the number of such masks that has to be used.

is beneficial for robustness. Nonetheless, so-called phase shift methods successfully use a set of patterns with sinusoidally varying intensities in one direction and constant intensity in the perpendicular direction (i.e., a more gradual stripe pattern than in the previous example). Each of the three sinusoidal patterns has the same amplitude but is 120° phase shifted with respect to each other. Intensity ratios in the images taken under each of the three patterns yield a unique position modulo the periodicity of the patterns. The sine patterns sum up to a constant intensity, so adding the three images yields the scene texture. The three subsequent projections yield dense range values plus texture. An example result is shown in Figure 1.6. These 3D measurements have been obtained with a system that works in real time (30 Hz depth + texture).

One can also design more intricate patterns that contain local spatial codes to identify parts of the projection pattern. An example is shown in Figure 1.7. The figure shows a face on which the single, checkerboard kind of pattern on the left is projected. The pattern is such that each column has its own distinctive signature. It consists of combinations of little white or black squares at the vertices of the checkerboard squares. 3D reconstructions obtained with this technique are shown in Figure 1.8. The use of this pattern only requires the acquisition of a single image. Hence, continuous projection

Fig. 1.6 3D results obtained with a phase-shift system. *Left*: 3D reconstruction without texture. *Right*: The same 3D reconstruction with texture, obtained by summing the three images acquired with the phase-shifted sine projections.

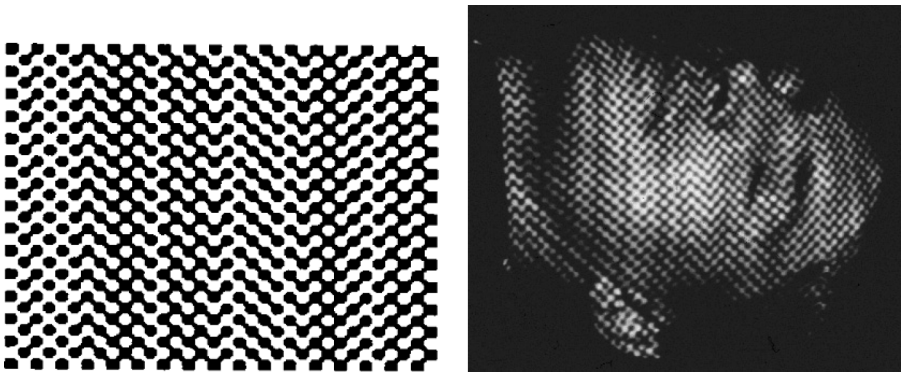

Fig. 1.7 Example of one-shot active range technique. *Left*: The projection pattern allowing disambiguation of its different vertical columns. *Right*: The pattern is projected on a face.

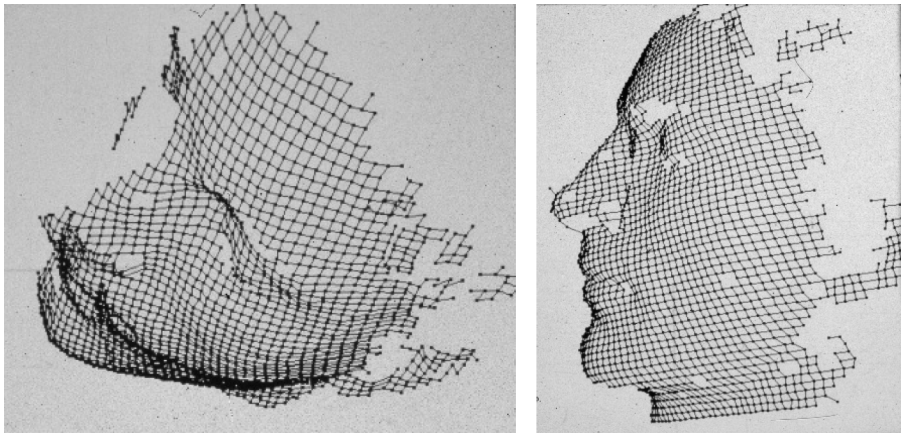

Fig. 1.8 Two views of the 3D description obtained with the active method of Figure 1.7.

in combination with video input yields a 4D acquisition device that can capture 3D shape (but not texture) and its changes over time. All these approaches with specially shaped projected patterns are commonly referred to as *structured light* techniques.

1.4 Other Methods

With the exception of time-of-flight techniques, all other methods in the taxonomy of Figure 1.1 are of less practical importance (yet). Hence,

only time-of-flight is discussed to a somewhat greater length. For the other approaches, only their general principles are outlined.

1.4.1 Time-of-Flight

The basic principle of time-of-flight sensors is the measurement of the duration before a sent out time-modulated signal — usually light from a laser — returns to the sensor. This time is proportional to the distance from the object. This is an active, single-vantage approach. Depending on the type of waves used, one calls such devices *radar* (electromagnetic waves of low frequency), *sonar* (acoustic waves), or *optical radar* (optical electromagnetic waves, including near-infrared).

A first category uses pulsed waves and measures the delay between the transmitted and the received pulse. These are the most often used type. A second category is used for smaller distances and measures phase shifts between outgoing and returning sinusoidal waves. The low level of the returning signal and the high bandwidth required for detection put pressure on the signal to noise ratios that can be achieved. Measurement problems and health hazards with lasers can be alleviated by the use of ultrasound. The bundle has a much larger opening angle then, and resolution decreases (a lot).

Mainly optical signal-based systems (typically working in the near-infrared) represent serious competition for the methods mentioned before. Such systems are often referred to as LIDAR (LIght Detection And Ranging) or LADAR (LAser Detection And Ranging, a term more often used by the military, where wavelengths tend to be longer, like 1,550 nm in order to be invisible in night goggles). As these systems capture 3D data point-by-point, they need to scan. Typically a horizontal motion of the scanning head is combined with a faster vertical flip of an internal mirror. Scanning can be a rather slow process, even if at the time of writing there were already LIDAR systems on the market that can measure 50,000 points per second. On the other hand, LIDAR gives excellent precision at larger distances in comparison to passive techniques, which start to suffer from limitations in image resolution. Typically, errors at tens of meters will be within a range of a few centimeters. Triangulation-based techniques require quite some baseline to achieve such small margins. A disadvantage is that surface

texture is not captured and that errors will be substantially larger for dark surfaces, which reflect little of the incoming signal. Missing texture can be resolved by adding a camera, as close as possible to the LIDAR scanning head. But of course, even then the texture is not taken from exactly the same vantage point. The output is typically delivered as a massive, unordered point cloud, which may cause problems for further processing. Moreover, LIDAR systems tend to be expensive.

More recently, 3D cameras have entered the market, that use the same kind of time-of-flight principle, but that acquire an entire 3D image at the same time. These cameras have been designed to yield real-time 3D measurements of smaller scenes, typically up to a couple of meters. So far, resolutions are still limited (in the order of 150×150 range values) and depth resolutions only moderate (couple of millimeters under ideal circumstances but worse otherwise), but this technology is making advances fast. It is expected that the price of such cameras will drop sharply soon, as some games console manufacturer's plan to offer such cameras as input devices.

1.4.2 Shape-from-Shading and Photometric Stereo

We now discuss the remaining, active techniques in the taxonomy of Figure 1.1.

'Shape-from-shading' techniques typically handle smooth, untextured surfaces. Without the use of structured light or time-of-flight methods these are difficult to handle. Passive methods like stereo may find it difficult to extract the necessary correspondences. Yet, people can estimate the overall shape quite well (qualitatively), even from a single image and under uncontrolled lighting. This would win it a place among the passive methods. No computer algorithm today can achieve such performance, however. Yet, progress has been made under simplifying conditions. One can use directional lighting with known direction and intensity. Hence, we have placed the method in the 'active' family for now. Gray levels of object surface patches then convey information on their 3D orientation. This process not only requires information on the sensor-illumination configuration, but also on the reflection characteristics of the surface. The complex relationship between gray levels

and surface orientation can theoretically be calculated in some cases — e.g., when the surface reflectance is known to be Lambertian — but is usually derived from experiments and then stored in 'reflectance maps' for table-lookup. For a Lambertian surface with known albedo and for a known light source intensity, the angle between the surface normal and the incident light direction can be derived. This yields surface normals that lie on a cone about the light direction. Hence, even in this simple case, the normal of a patch cannot be derived uniquely from its intensity. Therefore, information from different patches is combined through extra assumptions on surface smoothness. Neighboring patches can be expected to have similar normals. Moreover, for a smooth surface the normals at the visible rim of the object can be determined from their tangents in the image if the camera settings are known. Indeed, the 3D normals are perpendicular to the plane formed by the projection ray at these points and the local tangents to the boundary in the image. This yields strong boundary conditions. Estimating the lighting conditions is sometimes made part of the problem. This may be very useful, as in cases where the light source is the sun. The light is also not always assumed to be coming from a single direction. For instance, some lighting models consist of both a directional component and a homogeneous ambient component, where light is coming from all directions in equal amounts. Surface interreflections are a complication which these techniques so far cannot handle.

The need to combine normal information from different patches can be reduced by using different light sources with different positions. The light sources are activated one after the other. The subsequent observed intensities for the surface patches yield only a single possible normal orientation (not withstanding noise in the intensity measurements). For a Lambertian surface, three different lighting directions suffice to eliminate uncertainties about the normal direction. The three cones intersect in a single line, which is the sought patch normal. Of course, it still is a good idea to further improve the results, e.g., via smoothness assumptions. Such 'photometric stereo' approach is more stable than shape-from-shading, but it requires a more controlled acquisition environment. An example is shown in Figure 1.9. It shows a dome with 260 LEDs that is easy to assemble and disassemble (modular design, fitting

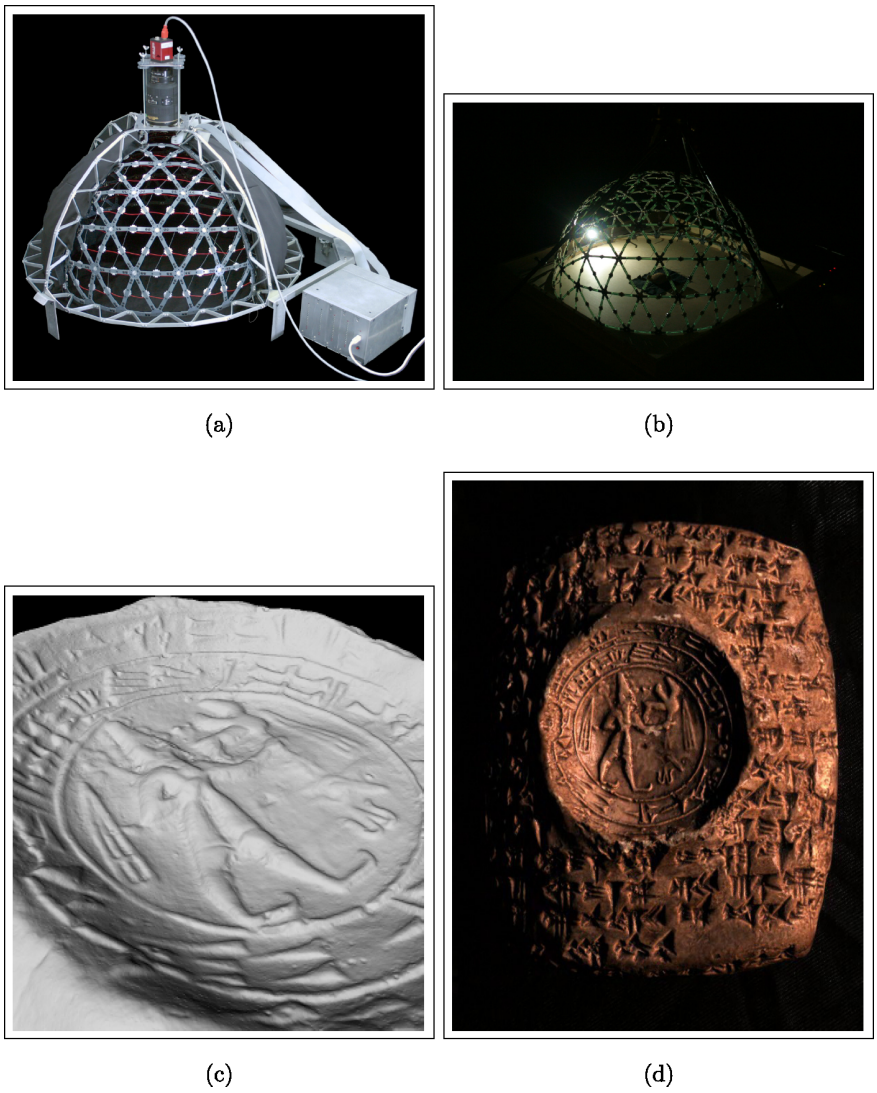

Fig. 1.9 (a) Mini-dome with different LED light sources, (b) scene with one of the LEDs activated, (c) 3D reconstruction of a cuneiform tablet, without texture, and (d) the same tablet with texture.

in a standard aircraft suitcase; see part (a) of the figure). The LEDs are automatically activated in a predefined sequence. There is one overhead camera. The resulting 3D reconstruction of a cuneiform tablet is shown in Figure 1.9(c) without texture, and in (d) with texture.

As with structured light techniques, one can try to reduce the number of images that have to be taken, by giving the light sources different colors. The resulting mix of colors at a surface patch yields direct information about the surface normal. In case 3 projections suffice, one can exploit the R-G-B channels of a normal color camera. It is like taking three intensity images in parallel, one per spectral band of the camera.

Note that none of the above techniques yield absolute depths, but rather surface normal directions. These can be integrated into full 3D models of shapes.

1.4.3 Shape-from-Texture and Shape-from-Contour

Passive single vantage methods include shape-from-texture and shape-from-contour. These methods do not yield true range data, but, as in the case of shape-from-shading, only surface orientation.

Shape-from-texture assumes that a surface is covered by a homogeneous texture (i.e., a surface pattern with some statistical or geometric regularity). Local inhomogeneities of the imaged texture (e.g., anisotropy in the statistics of edge orientations for an isotropic texture, or deviations from assumed periodicity) are regarded as the result of projection. Surface orientations which allow the original texture to be maximally isotropic or periodic are selected. Figure 1.10 shows an

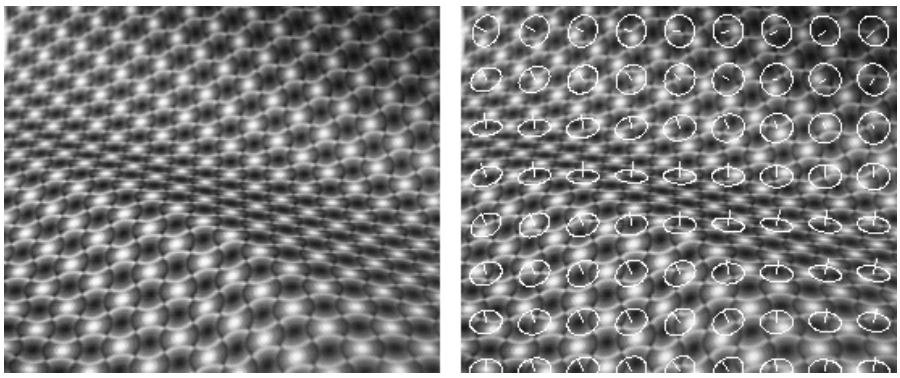

Fig. 1.10 *Left*: The regular texture yields a clear perception of a curved surface. *Right*: The result of a shape-from-texture algorithm.

example of a textured scene. The impression of an undulating surface is immediate. The right-hand side of the figure shows the results for a shape-from-texture algorithm that uses the regularity of the pattern for the estimation of the local surface orientation. Actually, what is assumed here is a square shape of the pattern's period (i.e., a kind of discrete isotropy). This assumption suffices to calculate the local surface orientation. The ellipses represent circles with such calculated orientation of the local surface patch. The small stick at their center shows the computed normal to the surface.

Shape-from-contour makes similar assumptions about the true shape of, usually planar, objects. Observing an ellipse, the assumption can be made that it actually is a circle, and the slant and tilt angles of the plane can be determined. For instance, in the shape-from-texture figure we have visualized the local surface orientation via ellipses. This 3D impression is compelling, because we tend to interpret the elliptical shapes as projections of what in reality are circles. This is an example of shape-from-contour as applied by our brain. The circle–ellipse relation is just a particular example, and more general principles have been elaborated in the literature. An example is the maximization of area over perimeter squared, as a measure of shape compactness, over all possible deprojections, i.e., surface patch orientations. Returning to our example, an ellipse would be deprojected to a circle for this measure, consistent with human vision. Similarly, symmetries in the original shape will get lost under projection. Choosing the slant and tilt angles that maximally restore symmetry is another example of a criterion for determining the normal to the shape. As a matter of fact, the circle–ellipse case also is an illustration for this measure. Regular figures with at least a 3-fold rotational symmetry yield a single orientation that could make up for the deformation in the image, except for the mirror reversal with respect to the image plane (assuming that perspective distortions are too small to be picked up). This is but a special case of the more general result, that a unique orientation (up to mirror reflection) also results when two copies of a shape are observed in the same plane (with the exception where their orientation differs by 0° or 180° in which case nothing can be said on the mere assumption that both shapes are identical). Both cases are more restrictive than

skewed mirror symmetry (without perspective effects), which yields a one-parameter family of solutions only.

1.4.4 Shape-from-Defocus

Cameras have a limited depth-of-field. Only points at a particular distance will be imaged with a sharp projection in the image plane. Although often a nuisance, this effect can also be exploited because it yields information on the distance to the camera. The level of defocus has already been used to create depth maps. As points can be blurred because they are closer or farther from the camera than at the position of focus, shape-from-defocus methods will usually combine more than a single image, taken from the same position but with different focal lengths. This should disambiguate the depth.

1.4.5 Shape-from-Silhouettes

Shape-from-silhouettes is a passive, multi-vantage approach. Suppose that an object stands on a turntable. At regular rotational intervals an image is taken. In each of the images, the silhouette of the object is determined. Initially, one has a virtual lump of clay, larger than the object and fully containing it. From each camera orientation, the silhouette forms a cone of projection rays, for which the intersection with this virtual lump is calculated. The result of all these intersections yields an approximate shape, a so-called visual hull. Figure 1.11 illustrates the process.

One has to be careful that the silhouettes are extracted with good precision. A way to ease this process is by providing a simple background, like a homogeneous blue or green cloth ('blue keying' or 'green keying'). Once a part of the lump has been removed, it can never be retrieved in straightforward implementations of this idea. Therefore, more refined, probabilistic approaches have been proposed to fend off such dangers. Also, cavities that do not show up in any silhouette will not be removed. For instance, the eye sockets in a face will not be detected with such method and will remain filled up in the final model. This can be solved by also extracting stereo depth from neighboring

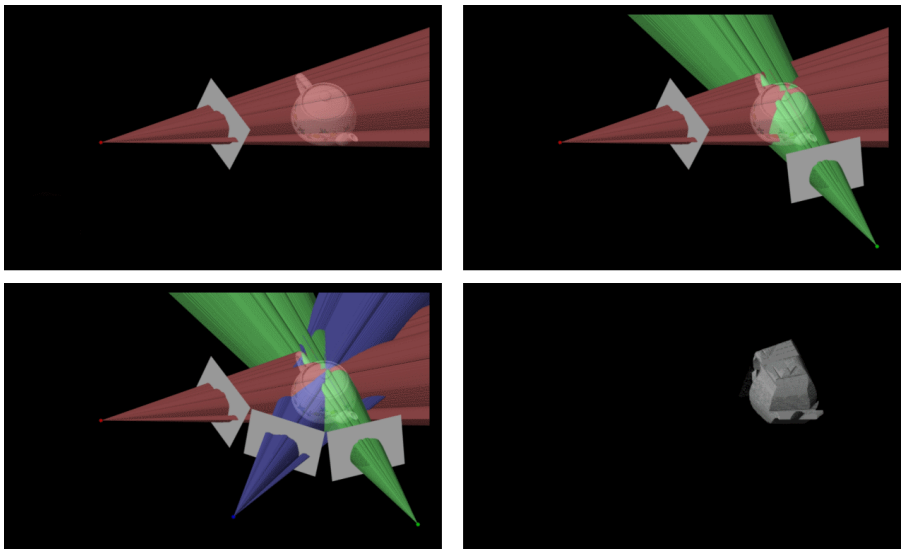

Fig. 1.11 The first three images show different backprojections from the silhouette of a teapot in three views. The intersection of these backprojections form the visual hull of the object, shown in the bottom right image. The more views are taken, the closer the visual hull approaches the true shape, but cavities not visible in the silhouettes are not retrieved.

viewpoints and by combining the 3D information coming from both methods.

The hardware needed is minimal, and very low-cost shape-from-silhouette systems can be produced. If multiple cameras are placed around the object, the images can be taken all at once and the capture time can be reduced. This will increase the price, and also the silhouette extraction may become more complicated. In the case video cameras are used, a dynamic scene like a moving person can be captured in 3D over time (but note that synchronization issues are introduced). An example is shown in Figure 1.12, where 15 video cameras were set up in an outdoor environment.

Of course, in order to extract precise cones for the intersection, the relative camera positions and their internal settings have to be known precisely. This can be achieved with the same self-calibration methods expounded in the following sections. Hence, also shape-from-silhouettes can benefit from the presented ideas and this is all the more interesting

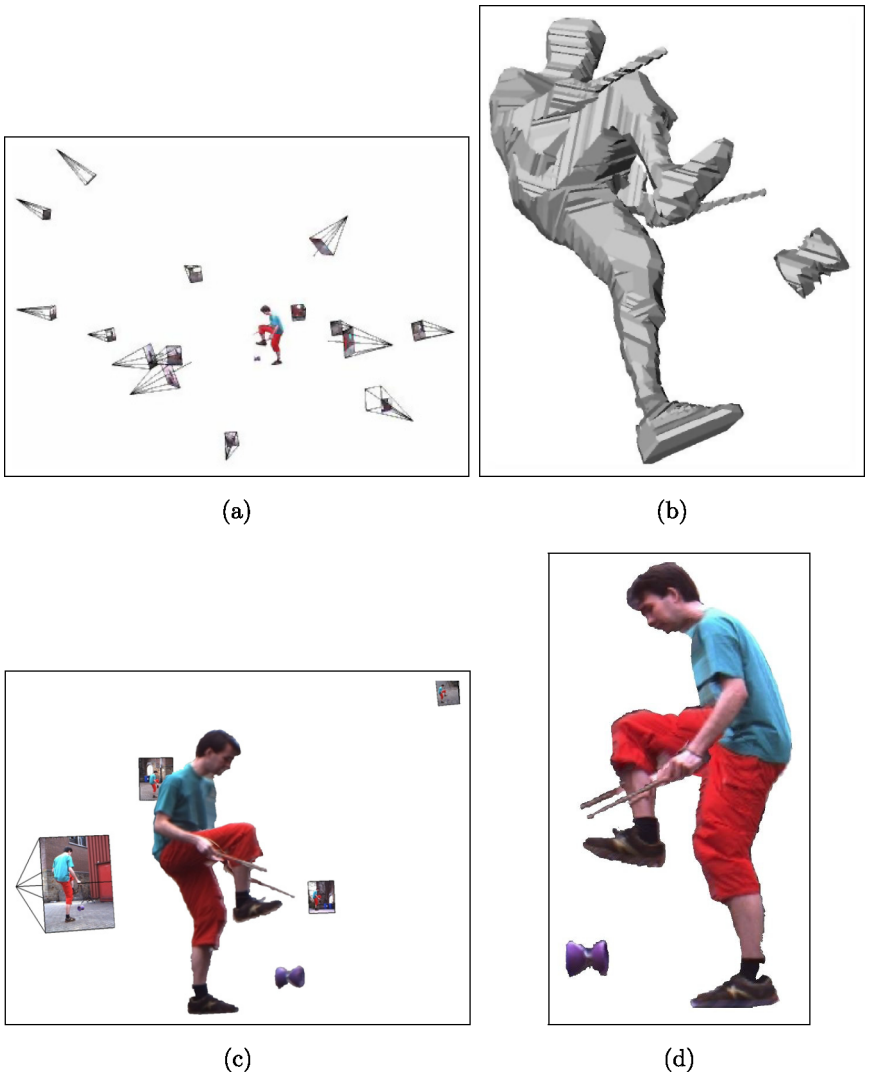

Fig. 1.12 (a) Fifteen cameras setup in an outdoor environment around a person,(b) a more detailed view on the visual hull at a specific moment of the action,(c) a detailed view on the visual hull textured by backprojecting the image colors, and (d) another view of the visual hull with backprojected colors. Note how part of the sock area has been erroneously carved away.

as this 3D extraction approach is among the most practically relevant ones for dynamic scenes ('motion capture').

1.4.6 Hybrid Techniques

The aforementioned techniques often have complementary strengths and weaknesses. Therefore, several systems try to exploit multiple techniques in conjunction. A typical example is the combination of shape-from-silhouettes with stereo as already hinted in the previous section. Both techniques are passive and use multiple cameras. The visual hull produced from the silhouettes provides a depth range in which stereo can try to refine the surfaces in between the rims, in particular at the cavities. Similarly, one can combine stereo with structured light. Rather than trying to generate a depth map from the images pure, one can project a random noise pattern, to make sure that there is enough texture. As still two cameras are used, the projected pattern does not have to be analyzed in detail. Local pattern correlations may suffice to solve the correspondence problem. One can project in the near-infrared, to simultaneously take color images and retrieve the surface texture without interference from the projected pattern. So far, the problem with this has often been the weaker contrast obtained in the near-infrared band. Many such integrated approaches can be thought of.

This said, there is no single 3D acquisition system to date that can handle all types of objects or surfaces. Transparent or glossy surfaces (e.g., glass, metals), fine structures (e.g., hair or wires), and too weak, too busy, to too repetitive surface textures (e.g., identical tiles on a wall) may cause problems, depending on the system that is being used. The next section discusses still existing challenges in a bit more detail.

1.5 Challenges

The production of 3D models has been a popular research topic already for a long time now, and important progress has indeed been made since the early days. Nonetheless, the research community is well aware of the fact that still much remains to be done. In this section we list some of these challenges.

As seen in the previous subsections, there is a wide variety of techniques for creating 3D models, but depending on the geometry and material characteristics of the object or scene, one technique may be much better suited than another. For example, untextured objects are

a nightmare for traditional stereo, but too much texture may interfere with the patterns of structured-light techniques. Hence, one would seem to need a battery of systems to deal with the variability of objects — e.g., in a museum — to be modeled. As a matter of fact, having to model the entire collections of diverse museums is a useful application area to think about, as it poses many of the pending challenges, often several at once. Another area is 3D city modeling, which has quickly grown in importance over the last years. It is another extreme in terms of conditions under which data have to be captured, in that cities represent an absolutely uncontrolled and large-scale environment. Also in that application area, many problems remain to be resolved.

Here is a list of remaining challenges, which we do not claim to be exhaustive:

- Many objects have an intricate shape, the scanning of which requires high precision combined with great agility of the scanner to capture narrow cavities and protrusions, deal with self-occlusions, fine carvings, etc.
- The types of objects and materials that potentially have to be handled — think of the museum example — are very diverse, like shiny metal coins, woven textiles, stone or wooden sculptures, ceramics, gems in jewellery and glass. No single technology can deal with all these surface types and for some of these types of artifacts there are no satisfactory techniques yet. Also, apart from the 3D shape the material characteristics may need to be captured as well.
- The objects to be scanned range from tiny ones like a needle to an entire construction or excavation site, landscape, or city. Ideally, one would handle this range of scales with the same techniques and similar protocols.
- For many applications, data collection may have to be undertaken on-site under potentially adverse conditions or implying transportation of equipment to remote or harsh environments.
- Objects are sometimes too fragile or valuable to be touched and need to be scanned 'hands-off'. The scanner needs to be

moved around the object, without it being touched, using portable systems.
- Masses of data often need to be captured, like in the museum collection or city modeling examples. More efficient data capture and model building are essential if this is to be practical.
- Those undertaking the digitization may or may not be technically trained. Not all applications are to be found in industry, and technically trained personnel may very well not be around. This raises the need for intelligent devices that ensure high-quality data through (semi-)automation, self-diagnosis, and effective guidance of the operator.
- In many application areas the money that can be spent is very limited and solutions therefore need to be relatively cheap.
- Also, precision is a moving target in many applications and as higher precisions are achieved, new applications present themselves that push for going even beyond. Analyzing the 3D surface of paintings to study brush strokes is a case in point.

These considerations about the particular conditions under which models may need to be produced, lead to a number of desirable, technological developments for 3D data acquisition:

- **Combined extraction of shape and surface reflectance.** Increasingly, 3D scanning technology is aimed at also extracting high-quality surface reflectance information. Yet, there still is an appreciable way to go before high-precision geometry can be combined with detailed surface characteristics like full-fledged BRDF (Bidirectional Reflectance Distribution Function) or BTF (Bidirectional Texture Function) information.
- **In-hand scanning.** The first truly portable scanning systems are already around. But the choice is still restricted, especially when also surface reflectance information is required and when the method ought to work with all types of materials, including metals, glass, etc. Also, transportable

here is supposed to mean more than 'can be dragged between places', i.e., rather the possibility to easily move the system around the object, ideally also by hand. But there also is the interesting alternative to take the objects to be scanned in one's hands, and to manipulate them such that all parts get exposed to the fixed scanner. This is not always a desirable option (e.g., in the case of very valuable or heavy pieces), but has the definite advantage of exploiting the human agility in presenting the object and in selecting optimal, additional views.

- **On-line scanning.** The physical action of scanning and the actual processing of the data often still are two separate steps. This may create problems in that the completeness and quality of the result can only be inspected after the scanning session is over and the data are analyzed and combined at the lab or the office. It may then be too late or too cumbersome to take corrective actions, like taking a few additional scans. It would be very desirable if the system would extract the 3D data on the fly, and would give immediate visual feedback. This should ideally include steps like the integration and remeshing of partial scans. This would also be a great help in planning where to take the next scan during scanning. A refinement can then still be performed off-line.
- **Opportunistic scanning.** Not a single 3D acquisition technique is currently able to produce 3D models of even a large majority of exhibits in a typical museum. Yet, they often have complementary strengths and weaknesses. Untextured surfaces are a nightmare for passive techniques, but may be ideal for structured light approaches. Ideally, scanners would automatically adapt their strategy to the object at hand, based on characteristics like spectral reflectance, texture spatial frequency, surface smoothness, glossiness, etc. One strategy would be to build a single scanner that can switch strategy on-the-fly. Such a scanner may consist of multiple cameras and projection devices, and by today's technology could still be small and light-weight.

- **Multi-modal scanning.** Scanning may not only combine geometry and visual characteristics. Additional features like non-visible wavelengths (UV,(N)IR) could have to be captured, as well as haptic impressions. The latter would then also allow for a full replay to the public, where audiences can hold even the most precious objects virtually in their hands, and explore them with all their senses.
- **Semantic 3D.** Gradually computer vision is getting at a point where scene understanding becomes feasible. Out of 2D images, objects and scene types can be recognized. This will in turn have a drastic effect on the way in which 'low'-level processes can be carried out. If high-level, semantic interpretations can be fed back into 'low'-level processes like motion and depth extraction, these can benefit greatly. This strategy ties in with the opportunistic scanning idea. Recognizing what it is that is to be reconstructed in 3D (e.g., a car and its parts) can help a system to decide how best to go about, resulting in increased speed, robustness, and accuracy. It can provide strong priors about the expected shape and surface characteristics.
- **Off-the-shelf components.** In order to keep 3D modeling cheap, one would ideally construct the 3D reconstruction systems on the basis of off-the-shelf, consumer products. At least as much as possible. This does not only reduce the price, but also lets the systems surf on a wave of fast-evolving, mass-market products. For instance, the resolution of still, digital cameras is steadily on the increase, so a system based on such camera(s) can be upgraded to higher quality without much effort or investment. Moreover, as most users will be acquainted with such components, the learning curve to use the system is probably not as steep as with a totally novel, dedicated technology.

Obviously, once 3D data have been acquired, further processing steps are typically needed. These entail challenges of their own. Improvements in automatic remeshing and decimation are definitely

still possible. Also solving large 3D puzzles automatically, preferably exploiting shape in combination with texture information, would be something in high demand from several application areas. Level-of-detail (LoD) processing is another example. All these can also be expected to greatly benefit from a semantic understanding of the data. Surface curvature alone is a weak indicator of the importance of a shape feature in LoD processing. Knowing one is at the edge of a salient, functionally important structure may be a much better reason to keep it in at many scales.

1.6 Conclusions

Given the above considerations, the 3D reconstruction of shapes from multiple, uncalibrated images is one of the most promising 3D acquisition techniques. In terms of our taxonomy of techniques, self-calibrating structure-from-motion is a passive, multi-vantage point strategy. It offers high degrees of flexibility in that one can freely move a camera around an object or scene. The camera can be hand-held. Most people have a camera and know how to use it. Objects or scenes can be small or large, assuming that the optics and the amount of camera motion are appropriate. These methods also give direct access to both shape and surface reflectance information, where both can be aligned without special alignment techniques. Efficient implementations of several subparts of such Structure-from-Motion pipelines have been proposed lately, so that the on-line application of such methods is gradually becoming a reality. Also, the required hardware is minimal, and in many cases consumer type cameras will suffice. This keeps prices for data capture relatively low.

References

[1] J. Y. Bouguet, "Camera calibration toolbox for matlab," http://www.vision.caltech.edu/bouguetj/calib_doc/.

[2] O. Faugeras, "What can be seen in three dimensions with an uncalibrated stereo rig," in *Computer Vision — (ECCV'92)*, pp. 563–578, vol. LNCS 588, Berlin/Heidelberg/New York/Tokyo: Springer-Verlag, 1992.

[3] G. H. Golub and C. F. V. Loan, *Matrix Computations*. Baltimore, ML, USA: The John Hopkins University Press, 1996.

[4] R. Hartley, "Estimation of relative camera positions for uncalibrated cameras," in *Computer Vision — (ECCV'92)*, pp. 579–587, vol. LNCS 588, Berlin/Heidelberg/New York/Tokyo: Springer-Verlag, 1992.

[5] R. Hartley, "Self-calibration from multiple views with a rotating camera," in *Computer Vision — (ECCV'94)*, pp. 471–478, vol. LNCS 800/801, Berlin/Heidelberg/New York/Tokyo: Springer-Verlag, 1994.

[6] R. Hartley, "Cheirality," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 41–61, 1998.

[7] R. Hartley and S. B. Kang, "Parameter-free radial distortion correction with center of distortion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1309–1321, doi:10.1109/TPAMI.2007.1147, June 2007.

[8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.

[9] F. Kahl, B. Triggs, and K. Åström, "Critical motions for auto-calibration when some intrinsic parameters can vary," *Journal of Mathematical Imaging and Vision*, vol. 13, no. 2, pp. 131–146, October 2000.

[10] H. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 10, pp. 133–135, 1981.

[11] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, 1981.

[12] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–777, June 2004.

[13] J. Philip, "A non-iterative algorithm for determining all essential matrices corresponding to five point Pairs," *The Photogrammetric Record*, vol. 15, no. 88, pp. 589–599, 1996.

[14] P. Sturm, "Critical motion sequences and conjugacy of ambiguous euclidean reconstructions," in *Proceedings of the 10th Scandinavian Conference on Image Analysis, Lappeenranta, Finland*, vol. I, (M. Frydrych, J. Parkkinen, and A. Visa, eds.), pp. 439–446, June 1997.

[15] P. Sturm, "Critical motion sequences for the self-calibration of cameras and stereo systems with variable focal length," in *British Machine Vision Conference, Nottingham, England*, pp. 63–72, September 1999.

[16] B. Triggs, "Autocalibration and the absolute quadric," in *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 609–614, Washington, DC, USA: IEEE Computer Society, 1997.

[17] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *Radiometry*, pp. 221–244, 1992.

[18] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *ICCV*, pp. 666–673, 1999.