1

3D Face Modeling

Boulbaba Ben Amor,[1] Mohsen Ardabilian,[2] and Liming Chen[2]
[1]*Institut Mines-Télécom/Télécom Lille 1, France*
[2]*Ecole Centrale de Lyon, France*

Acquiring, modeling, and synthesizing realistic 3D human faces and their dynamics have emerged as an active research topic in the border area between the computer vision and computer graphics fields of research. This has resulted in a plethora of different acquisition systems and processing pipelines that share many fundamental concepts as well as specific implementation details. The research community has investigated the possibility of targeting either end-to-end consumer-level or professional-level applications, such as facial geometry acquisition for 3D-based biometrics and its dynamics capturing for expression cloning or performance capture and, more recently, for 4D expression analysis and recognition. Despite the rich literature, reproducing realistic human faces remains a distant goal because the challenges that face 3D face modeling are still open. These challenges include the motion speed of the face when conveying expressions, the variabilities in lighting conditions, and pose. In addition, human beings are very sensitive to facial appearance and quickly sense any anomalies in 3D geometry or dynamics of faces. The techniques developed in this field attempt to recover facial 3D shapes from camera(s) and reproduce their actions. Consequently, they seek to answer the following questions:

- How can one recover the facial shapes under pose and illumination variations?
- How can one synthesize realistic dynamics from the obtained 3D shape sequences?

This chapter provides a brief overview of the most successful existing methods in the literature by first introducing basics and background material essential to understand them. To this end, instead of the *classical* passive/active taxonomy of 3D reconstruction techniques, we propose here to categorize approaches according to whether they are able to acquire faces in action or they can only capture them in a static state. Thus, this chapter is preliminary to

3D Face Modeling, Analysis and Recognition, First Edition.
Edited by Mohamed Daoudi, Anuj Srivastava and Remco Veltkamp.

the following chapters that use static or dynamic facial data for face analysis, recognition, and expression recognition.

1.1 Challenges and Taxonomy of Techniques

Capturing and processing human geometry is at the core of several applications. To work on 3D faces, one must first be able to recover their shapes. In the literature, several acquisition techniques exist that are either dedicated to specific objects or are general. Usually accompanied by geometric modeling tools and post-processing of 3D entities (3D point clouds, 3D mesh, volume, etc.), these techniques provide complete solutions for 3D full object reconstruction. The acquisition quality is mainly linked to the accuracy of recovering the z-coordinate (called depth information). It is characterized by loyalty reconstruction, in other words, by data quality, the density of 3D face models, details preservation (regions showing changes in shapes), etc. Other important criteria are the acquisition time, the ease of use, and the sensor's cost. In what follows, we report the main *extrinsic* and *intrinsic* factors which could influence the modeling process.

- ☞ *Extrinsic factors*. They are related to the environmental conditions of the acquisition and the face itself. In fact, human faces are globally similar in terms of the position of main features (eyes, mouth, nose, etc.), but can vary considerably in details across (i) their variabilities due to facial deformations (caused by expressions and mouth opening), subject aging (wrinkles), etc, and (ii) their specific details as skin color, scar tissue, face asymmetry, etc. The environmental factors refer to lighting conditions (controlled or ambient) and changes in head pose.
- ☞ *Intrinsic factors*. They include sensor cost, its intrusiveness, manner of sensor use (cooperative or not), spatial and/or temporal resolutions, measurement accuracy and the acquisition time, which allows us to capture moving faces or simply faces in static state.

These challenges arise when acquiring static faces as well as when dealing with faces in action. Different applications have different requirements. For instance, in the computer graphics community, the results of performance capture should exhibit a great deal of spatial fidelity and temporal accuracy to be an authentic reproduction of a real actors' performance. Facial recognition systems, on the other hand, require the accurate capture of person-specific details. The movie industry, for instance, may afford a 3D modeling pipeline system with special purpose hardware and highly specialized sensors that require manual calibration. When deploying a 3D acquisition system for facial recognition at airports and in train stations, however, cost, intrusiveness, and the need of user cooperation, among others, are important factors to consider. In ambient intelligence applications where a user-specific interface is required, facial expression recognition from 3D sequences emerges as a research trend instead of 2D-based techniques, which are sensitive to changes and pose variations. Here, also, sensor cost and its capability to capture facial dynamics are important issues. Figure 1.1 shows a new 3D face modeling-guided taxonomy of existing reconstruction approaches. This taxonomy proposes two categories: The first category targets 3D static face modeling, while the approaches belonging to the second category try to capture facial shapes in action (i.e., in 3D+t domain). In the level below, one finds different approaches based on concepts presented

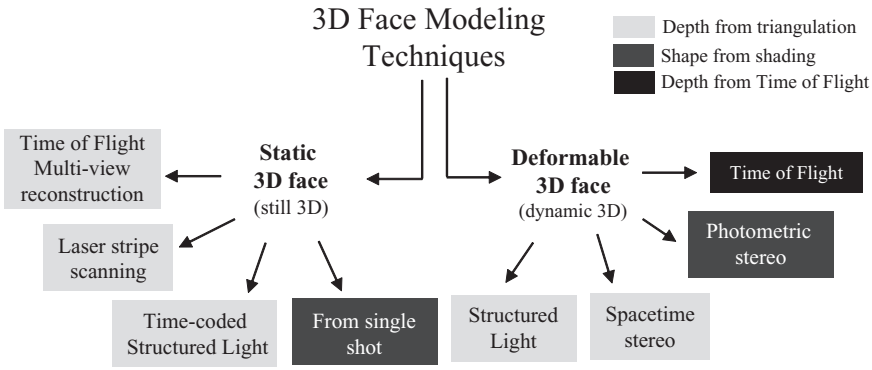

Figure 1.1 Taxonomy of 3D face modeling techniques

in section 1.2. In static face category, the multi-view stereo reconstruction uses the *optical triangulation* principle to recover the depth information of a scene from two or more projections (images). The same mechanism is unconsciously used by our brain to work out how far an object is. The correspondence problem in multi-view approaches is solved by looking for pixels that have the same appearance in the set of images. This is known as *stereo-matching* problem. Laser scanners use the *optical triangulation* principle, this time called *active* by replacing one camera with a laser source that emits a stripe in the direction of the object to scan. A second camera from a different viewpoint captures the projected pattern. In addition to one or several cameras, time-coded structured-light techniques use a light source to project on the scene a set of light patterns that are used as *codes* for finding correspondences between stereo images. Thus, they are also based on the *optical triangulation* principle.

The moving face modeling category, unlike the first one, needs fast processing for 3D shape recovery, thus, it tolerates scene motion. The structured-light techniques using one complex pattern is one solution. In the same direction, the work called *Spacetime faces* shows remarkable results in dynamic 3D shape modeling, by employing random colored light on the face to solve the stereo matching problem. Time-of-flight-based techniques could be used to recover the dynamic of human body parts such as the faces but with a modest shape accuracy. Recently, photometric stereo has been used to acquire 3D faces because it can recover a dense normal field from a surface. In the following sections, this chapter first gives basic principles shared by the techniques mentioned earlier, then addresses the details of each method.

1.2 Background

In the projective pinhole camera model, a point P in the 3D space is imaged into a point p on the image plane. p is related to P with the following formula:

$$p = MP = KR[I|t]P, \tag{1.1}$$

where p and P are represented in homogeneous coordinates, M is a 3×4 projection matrix, and I is the 3×3 identity matrix. M can be decomposed into two components: the intrinsic

parameters and the extrinsic parameters. Intrinsic parameters relate to the internal parameters of the camera, such as the image coordinates of the principal point, the focal length, pixel shape (its aspect ratio), and the skew. They are represented by the 3×3 upper triangular matrix K. Extrinsic (or external) parameters relate to the pose of the camera, defined by the 3×3 rotation matrix R and its position t with respect to a global coordinate system. *Camera calibration* is the process of estimating the intrinsic and extrinsic parameters of the cameras.

3D reconstruction can be roughly defined as the inverse of the imaging process; given a pixel p on one image, 3D reconstruction seeks to find the 3D coordinates of the point P that is imaged onto p. This is an ill-posed problem because with the inverse imaging process a pixel p maps into a ray v that starts from the camera center and passes through the pixel p. The ray direction $\vec{v}$ can be computed from the camera pose R and its intrinsic parameters K as follows;

$$\vec{v} = \frac{R^{-1}K^{-1}p}{\|R^{-1}K^{-1}p\|} \tag{1.2}$$

1.2.1 Depth from Triangulation

If q is the image of the same 3D point P taken by another camera from a different viewing angle, then the 3D coordinates of P can be recovered by estimating the intersection of the two rays, v_1 and v_2, that start from the camera centers passing, respectively, through p and q. This is known as the *optical triangulation* principle. p and q are called *corresponding* or *matching* pixels because they are the images of the same 3D point P.

A 3D point P is the intersection of $n (n > 1)$ rays v_i passing through the optical centers c_i of cameras $\{C_i\}$ where $i = 1, \ldots, n$. This can also be referred to *passive optical triangulation*. As illustrated in Figure 1.2, all points on v_i project to p_i, given a set of corresponding pixels p_i captured by the cameras C_i, and their corresponding rays v_i, the 3D location of P can be found by intersecting the rays v_i. In practice, however, these rays will often not intersect.

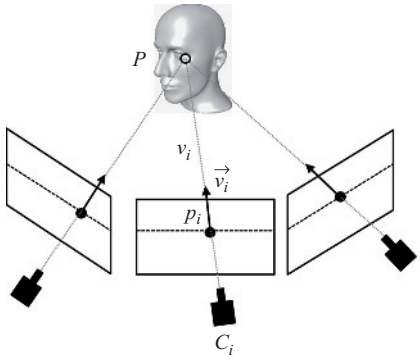

Figure 1.2 Multiview stereo determines the position of a point in space by finding the intersection of the rays v_i passing through the center of projection c_i of the ith camera and the projection of the point P in each image, p_i

Instead, we look for the optimal value of P that lies closest to the rays v_i. Mathematically, if K_i, R_i, t_i are the parameters of the camera C_i, where K_i is the 3×3 matrix that contains the intrinsic parameters of the camera and R_i and t_i are the pose of the ith camera with respect to the world coordinate system, the rays v_i originating at C_i and passing through p_i are in the direction of $R_i^{-1} K_i^{-1} p_i$. The optimal value of P that lies closest to all the rays $\vec{v}_i$, p minimizes the distance:

$$\|c_j + d_j \vec{v}_j - p\|^2 \tag{1.3}$$

Methods based on the *optical triangulation* need to solve two problems: *(i)* the matching problem, and *(ii)* the reconstruction problem. The correspondence problem consists of finding matching points across the different cameras. Given the corresponding points, the reconstruction problem consists of computing a 3D disparity map of the scene, which is equivalent to the depth map (z-coordinate on each pixel). Consequently, the quality of the reconstruction depends crucially on the solution to the correspondence problem. For further reading on *stereo vision* (cameras calibration, stereo matching algorithms, reconstruction, etc.), we refer the reader to download the PDF of the Richard Szeliski's *Computer Vision: Algorithms and Applications* available at http://szeliski.org.[1]

Existing *optical triangulation*-based 3D reconstruction techniques, such as multi-view stereo, structured-light techniques, and laser-based scanners, differ in the way the correspondence problem is solved. Multiview stereo reconstruction uses the triangulation principle to recover the depth map of a scene from two or more projections. The same mechanism is unconsciously used by our brain to work out how far an object is. The correspondence problem in stereo vision is solved by looking for pixels that have the same appearance in the set of images. This is known as *stereo matching*. Structured-light techniques use, in addition to camera(s), a light source to project on the scene a set of light patterns that are used as *codes* for finding correspondences between stereo images. Laser scanners use the triangulation principle by replacing one camera with a laser source that emits a laser ray in the direction of the object to scan. A camera from a different viewpoint captures the projected pattern.

1.2.2 Shape from Shading

Artists have reproduced, in paintings, illusions of depth using lighting and shading. Shape From Shading (SFS) addresses the shape recovery problem from a gradual variation of shading in the image. Image formation is a key ingredient to solve the SFS problem. In the early 1970s, Horn was the first to formulate the SFS problem as that of finding the solution of a nonlinear first-order Partial Differential Equation (PDE) also called the brightness equation. In the 1980s, the authors address the computational part of the problem, directly computing numerical solutions. Bruss and Brooks asked questions about the existence and uniqueness of solutions. According to the Lambertian model of image formation, the gray level at an image pixel depends on the light source direction and surface normal. Thus, the aim is to recover the illumination source

[1] http://szeliski.org/Book/

and the surface shape at each pixel. According to Horn's formulation of SFS problem, the brightness equation arises as:

$$I(x, y) = R(\vec{n}(x, y)), \tag{1.4}$$

where, (x, y) are the coordinates of a pixel; R, the reflectance map and I the brightness image. Usually, SFS approaches, particularly those dedicated to face shape recovery, adopt the Lambartian property of the surface. In which case, the reflectance map is the cosine of the angle between light vector $\vec{L}(x, y)$ and the normal vector $\vec{n}(x, y)$ to the surface:

$$R = \cos(\vec{L}, \vec{n}) = \frac{\vec{L}}{|\vec{L}|} \cdot \frac{\vec{n}}{|\vec{n}|}, \tag{1.5}$$

where R, $\vec{L}$ and $\vec{n}$ depends on (x, y). Since the first SFS technique developed by Horn, many different approaches have emerged; active SFS which requires calibration to simplify the solution finding has achieved impressive results.

1.2.3 Depth from Time of Flight (ToF)

Time of flight provides a direct way to acquire 3-D surface information of objects or scenes outputting 2.5 D, or depth, images with a real-time capability. The main idea is to estimate the time taken for the light projected by an illumination source to return from the scene or the object surface. This approach usually requires nano-second timing to resolve surface measurements to millimeter accuracy. The object or scene is actively illuminated with a nonvisible light source whose spectrum is usually nonvisible infrared, e.g. 780 nm. The intensity of the active signal is modulated by a cosine-shaped signal of frequency f. The light signal is assumed to have a constant speed, c, and is reflected by the scene or object surface. The distance d is estimated from the phase shift θ in radian between the emitted and the reflected signal, respectively:

$$d = \frac{c}{2f}\frac{\theta}{2\pi} \tag{1.6}$$

While conventional imaging sensors consists of multiple photo diodes, arranged within a matrix to provide an image of, e.g., color or gray values, a ToF sensor, for instance a photon mixing device (PMD) sensor, simultaneously acquires a distance value for each pixel in addition to the common intensity (gray) value. Compared with conventional imaging sensors, a PMD sensor is a standard CMOS sensor that benefits from these functional improvements. The chip includes all intelligence, which means that the distance is computed per pixel. In addition, some ToF cameras are equipped with a special pixel-integrated circuit, which guarantees the independence to sunlight influence by the suppression of background illumination (SBI).

1.3 Static 3D Face Modeling

1.3.1 Laser-stripe Scanning

Laser-stripe triangulation uses the well-known optical triangulation described in section 1.2. A laser line is swept across the object where a CCD array camera captures the reflected light, its shape gives the depth information. More formally, as illustrated in Figure 1.3, a slit laser beam, generated by a light projecting optical system, is projected on the object to be measured, and its reflected light is received by a CCD camera for triangulation. Then, 3D distance data for one line of slit light are obtained. By scanning slit light with a *galvanic mirror*, 3D data for the entire object to be measured are obtained. By measuring the angle $2\pi - \theta$, formed by the baseline d (distance between the light-receiving optical system and the light-projecting optical system) and by a laser beam to be projected, one can determine the z-coordinate by triangulation. The angle θ is determined by an instruction value of the galvanic mirror. Absolute coordinates for laser beam position on the surface of the object, denoted by p, are obtained from congruence conditions of triangles, by

$$\frac{z}{f_0} = \frac{d - z.\tan(\theta)}{p}. \tag{1.7}$$

This gives the z-coordinate, by

$$z = \frac{df_0}{p + f_0 \tan(\theta)}. \tag{1.8}$$

Solve question 1 in section 5.5.3 for the proof.

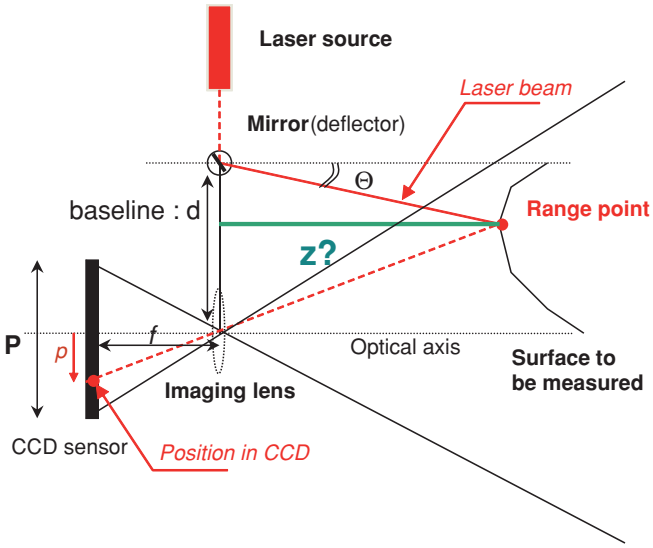

Figure 1.3 Optical triangulation geometry for a laser-stripe based scanner

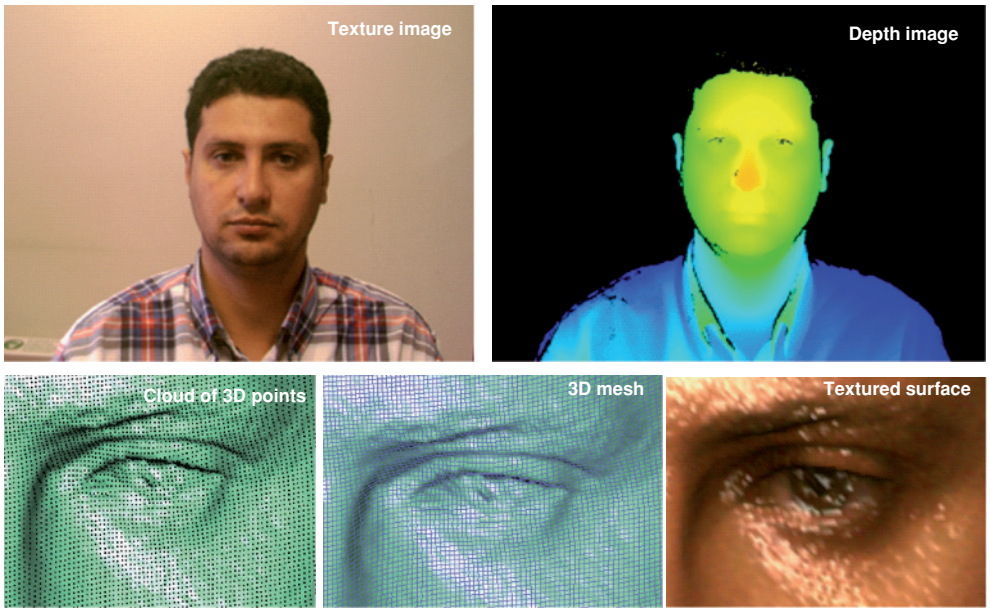

Figure 1.4 One example of 3D face acquisition based on laser stripe scanning (using Minolta VIVID 910). Different representations are given, from the left: texture image, depth image, cloud of 3D points, 3D mesh, and textured shape

The Charged Couple Device (CCD) is the widely used light-receiving optical system to digitize the point laser image. CCD-based sensors avoid the beam spot reflection and stray light effects and provide more accuracy because of the single-pixel resolution. Another factor that affects the measurement accuracy is the difference in the surface characteristic of the measured object from the calibration surface. Usually calibration should be performed on similar surfaces to ensure measurement accuracy. Using laser as a light source, this method has proven to be able to provide measurement at a much higher depth range than other passive systems with good discrimination of noise factors. However, this line-by-line measurement technique is relatively slow. The laser-based techniques can give very accurate 3D information for a rigid body even with a large depth. However, this method is time consuming for real measurement since it obtains 3D geometry on a line at a time. The area scanning-based methods such as time-coded structured light (see section 1.3.2) are certainly faster.

An example of acquired face using these technique is given by Figure 1.4. It illustrates the good quality of the reconstruction when office environment acquisition conditions are considered, the subject is distant of 1 m from the sensor and remains stable for a few seconds.

1.3.2 Time-coded Structured Light

The most widely used acquisition systems for face are based on structured light by virtue of reliability for recovering complex surface and accuracy. That consists in projecting a light pattern and imaging the illuminated object, a face for instance, from one or more points of

Figure 1.5 (a) Binary-coded patterns projection for 3D acquisition, (b) *n*-ary-coded coded patterns projection for 3D acquisition

view. Correspondences between image points and points of the projected pattern can be easily found. Finally the decoded points can be triangulated, and depth is recovered. The patterns are designed so that code words are assigned to a set of pixels.

A code word is assigned to a coded pixel to ensure a direct mapping from the code words to the corresponding coordinates of the pixel in the pattern. The code words are numbers and they are mapped in the pattern by using gray levels, color or geometrical representations. Pattern projection techniques can be classified according to their coding strategy: time-multiplexing, neighborhood codification, and direct codification. Time-multiplexing consists in projecting code words as sequence of patterns along time, so the structure of every pattern can be very simple. In spite of increased complexity, neighborhood codification represents the code words in a unique pattern. Finally, direct codification defines a code word for every pixel; equal to the pixel gray level or color.

One of the most commonly exploited strategies is based on temporal coding. In this case, a set of patterns are successively projected onto the measuring surface. The code word for a given pixel is usually formed by the sequence of illumination values for that pixel across the projected patterns. Thus, the codification is called *temporal* because the bits of the code words are multiplexed in time. This kind of pattern can achieve high accuracy in the measurements. This is due to two factors: First, because multiple patterns are projected, the code word basis tends to be small (usually binary) and hence a small set of primitives is used, being easily distinguishable among each other. Second, a coarse-to-fine paradigm is followed, because the position of a pixel is encoded more precisely while the patterns are successively projected.

During the three last decades, several techniques based on time-multiplexing have appeared. These techniques can be classified into three categories: binary codes (Figure 1.5**a**), *n*-ary codes (Fig. 1.5**b**), and phase-shifting techniques.

- **Binary codes.** In binary code, only two illumination levels are used. They are coded as 0 and 1. Each pixel of the pattern has its code word formed by the sequence of 0 and 1 corresponding to its value in every projected pattern. A code word is obtained once the sequence is completed. In practice, illumination source and camera are assumed to be strongly calibrated and hence only one of both pattern axes is encoded. Consequently, black and white strips are used to compose patterns – black corresponding to 0 and white 1, m patterns encode 2^m stripes. The maximum number of patterns that can be projected is the resolution in pixels of the projector device; however, because the camera cannot always perceive such narrow stripes, reaching this value is not recommended. It should be noticed that all pixels belonging to a similar stripe in the highest frequency pattern share the same

code word. Therefore, before triangulating, it is necessary to calculate either the center of every stripe or the edge between two consecutive stripes. The latter has been shown to be the best choice.

- ***N*-ary codes.** The main drawback of binary codes is the large number of patterns to be projected. However, the fact that only two intensities are projected eases the segmentation of the imaged patterns. The number of patterns can be reduced by increasing the number of intensity levels used to encode the stripes. A first mean is to use multilevel Gray code based on color. This extension of Gray code is based on an alphabet of n symbols; each symbol is associated with a certain RGB color. This extended alphabet makes it possible to reduce the number of patterns. For instance, with binary Gray code, m patterns are necessary to encode 2^m stripes. With an n-ary code, n^m stripes can be coded using the same number of patterns.
- **Phase shifting.** Phase shifting is a well-know principle in the pattern projection approach for 3D surface acquisition. Here, a set of sinusoidal patterns is used. The intensities of a pixel $p(x, y)$ in each pattern is given by:

$$\begin{aligned} I_1(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\, cos\,(\phi(x, y) - \theta), \\ I_2(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\, cos\,(\phi(x, y)), \\ I_3(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\, cos\,(\phi(x, y) + \theta). \end{aligned} \tag{1.9}$$

$I_0(x, y)$ is the background or the texture information, $I_{\text{mod}}(x, y)$ is the signal modulation amplitude, and $I_1(x, y)$, $I_2(x, y)$ and $I_3(x, y)$ are the intensities of the three patterns. $\phi(x, y)$ is the phase value and $\theta = \frac{2\pi}{3}$ is a constant. Three images of the object are used to estimate a wrapped phase value $\hat{\phi}(x, y)$ by:

$$\hat{\phi}(x, y) = \arctan\left\{\sqrt{3}\,\frac{I_1(x, y) - I_3(x, y)}{2\,I_2(x, y) - I_1(x, y) - I_3(x, y)}\right\} \tag{1.10}$$

The wrapped phase is periodic and needs to be unwrapped to obtain an absolute phase value $\phi'(x, y) = \phi(x, y) + 2k\pi$, where k is an integer representing the period or the number of the fringe. Finally the 3D information is recovered based on the projector-camera system configuration. Other pattern configurations of these patterns have been proposed. For instance, Zhang and Yau proposed a real-time 3D shape measurement based on a modified three-step phase-shifting technique (Zhang et al., 2007) (Figure 1.6). They called the modified patterns 2+1 phase shifting approach. According to this approach, the patterns and phase estimation are given by

$$\begin{aligned} I_1(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\, cos\left(\phi(x, y) - \frac{\pi}{2}\right), \\ I_2(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\, cos\,(\phi(x, y)), \\ I_3(x, y) &= I_0(x, y), \end{aligned} \tag{1.11}$$

$$\hat{\phi}(x, y) = \arctan\left\{\frac{I_1(x, y) - I_3(x, y)}{I_2(x, y) - I_3(x, y)}\right\}. \tag{1.12}$$

Figure 1.6 The high-resolution and real-time 3D shape measurement system proposed by Zhang and Yau (2007) is based on the modified 2 + 1 phase-shifting algorithm and particularly adapted for face acquisition. The data acquisition speed is as high as 60 frames per second while the image resolution is 640 × 480 pixels per frame. Here a photograph captured during the experiment is illustrated. The left side of the image shows the subject, whereas the right side shows the real-time reconstructed geometry

A robust phase unwrapping approach called "multilevel quality-guided phase unwrapping algorithm" is also proposed in Zhang et al. (2007).

Ouji et al. (2011) proposed a cost-effective 3D video acquisition solution with a 3D super-resolution scheme, using three calibrated cameras coupled with a non-calibrated projector device, which is particularly suited to 3D face scanning, that is, rapid, easily movable, and robust to ambient lighting conditions. Their solution is a hybrid stereovision and phase-shifting approach that not only takes advantage of the assets of stereovision and structured light but also overcomes their weaknesses. First, a 3D sparse model is estimated from stereo matching with a fringe-based resolution and a sub-pixel precision. Then projector parameters are automatically estimated through an inline stage. A dense 3D model is recovered by the intrafringe phase estimation, from the two sinusoidal fringe images and a texture image, independently from the left, middle, and right cameras. Finally, the left, middle, and right 3D dense models are fused to produce the final 3D model, which constitutes a spatial super-resolution. In contrast with previous methods, camera-projector calibration and phase-unwrapping stages are avoided.

1.3.3 Multiview Static Reconstruction

The aim of multiview stereo (MVS) reconstruction is twofold. Firstly, it allows to reinforce constraints on stereo matching, discard false matches, and increase the precision of good matches. Secondly, spatial arrangement of cameras allows covering the entire face. To reduce the complexity, as well as achieve high quality reconstruction, multiview reconstruction approaches usually proceed in a coarse-to-fine sequence. Finally, multiview approaches involve high resolution images captured in real time, whereas the processing stage requires tens of minutes. MVS scene and object reconstruction approaches can be organized into four categories. The first category operates first by estimating a cost function on a 3D volume and then extracting

a surface from this volume. A simple example of this approach is the voxel-coloring algorithm and its variants (Seitz and Dyer, 1997; Treuille et al., 2004). The second category of approaches, based on voxels, level sets, and surface meshes, works by iteratively evolving a surface to decrease or minimize a cost function. For example, from an initial volume, space carving progressively removes inconsistent voxels. Other approaches represent the object as an evolving mesh (Hernandez and Schmitt, 2004; Yu et al., 2006) moving as a function of internal and external forces. In the third category are image-space methods that estimate a set of depth maps. To ensure a single consistent 3D object interpretation, they enforce consistency constraints between depth maps (Kolmogorov and Zabih, 2002; Gargallo and Sturm, 2005) or merge the set of depth maps into a 3D object as a post process (Narayanan et al., 1998). The final category groups approaches that first extract and matches a set of feature points. A surface is then fitted to the reconstructed features (Morris and Kanade, 2000; Taylor, 2003). Seitz et al. (2006) propose an excellent overview and categorization of MVS. 3D face reconstruction approaches use a combination of methods from these categories.

Furukawa and Ponce (2009) proposed a MVS algorithm that outputs accurate models with a fine surface. It implements multiview stereopsis as a simple match, expand, and filter procedure. In the matching step, a set of features localized by Harris operator and difference-of-Gaussians algorithms are matched across multiple views, giving a sparse set of patches associated with salient image regions. From these initial matches, the two next steps are repeated n times ($n = 3$ in experiments). In the expansion step, initial matches are spread to nearby pixels to obtain a dense set of patches. Finally in the filtering step, the visibility constraints are used to discard incorrect matches lying either in front of or behind the observed surface. The MVS approach proposed by Bradley et al. (2010) is based on an iterative binocular stereo method to reconstruct seven surface patches independently and to merge into a single high resolution mesh. At this stage, face details and surface texture help guide the stereo algorithm. First, depth maps are created from pairs of adjacent rectified viewpoints. Then the most prominent distortions between the views are compensated by a *scaled-window matching* technique. The resulted depth images are converted to 3D points and fused into a single dense point cloud. A triangular mesh from the initial point cloud is reconstructed over three steps: down-sampling, outliers removal, and triangle meshing. Sample reconstruction results of this approach are shown in Figure 1.7.

The 3D face acquisition approach proposed by Beeler et al. (2010), which is built on the survey paper, takes inspiration from Furukawa and Ponce (2010). The main difference lies in a refinement formulation. The starting point is the established approach for refining recovered 3D data on the basis of a data-driven photo-consistency term and a surface-smoothing term, which has been research topic. These approaches differ in the use of a second-order anisotropic formulation of the smoothing term, and we argue that it is particularly suited to faces. Camera calibration is achieved in a pre-processing stage.

The run-time system starts with a pyramidal pairwise stereo matching. Results from lower resolutions guide the matching at higher-resolutions. The face is first segmented based on cues of background subtraction and skin color. Images from each camera pair are rectified. An image pyramid is then generated by factor of two downsampling using Gaussian convolution and stopping at approximately 150×150 pixels for the lowest layer. Then a dense matching is established between pairwise neighboring cameras, and each layer of the pyramid is processed as follows: Matches are computed for all pixels on the basis of normalized cross correlation (NCC) over a square window (3×3). The disparity is computed to sub-pixel accuracy and used to constrain the search area in the following layer. For each pixel, smoothness, uniqueness,

Figure 1.7 Sample results on 3D modeling algorithm for calibrated multiview stereopsis proposed by Furukawa and Ponce (2010) that outputs a quasi-dense set of rectangular patches covering the surfaces visible in the input images. In each case, one of the input images is shown on the left, along with two views of textured-mapped reconstructed patches and shaded polygonal surfaces. Copyright © 2007, IEEE

and ordering constraints are checked, and the pixels that do not fulfill these criteria are reached using the disparity estimated at neighboring pixels. The limited search area ensures smoothness and ordering constraints, but the uniqueness constraint is enforced again by disparity map refinement. The refinement is defined as a linear combination of a photometric consistency term, d_p, and a surface consistency term, d_s, balanced both by a user-specified smoothness parameter, w_s, and a data-driven parameter, w_p, to ensure that the photometric term has the greatest weight in regions with good feature localization. d_p favors solutions with high NCC, whereas d_s favors smooth solutions. The refinement is performed on the disparity map and later on the surface. Both are implemented as iterative processes.

The refinement results in surface geometry that is smooth across skin pores and fine wrinkles because the disparity change across such a feature is too small to detect. The result is flatness and lack of realism in synthesized views of the face. On the other hand, visual inspection shows the obvious presence of pores and fine wrinkles in the images. This is due to the fact that light reflected by a diffuse surface is related to the integral of the incoming light. In small concavities, such as pores, part of the incoming light is blocked and the point thus appears darker. This has been exploited by various authors (e.g., Glencross et al., 2008)) to infer local geometry variation. In this section, we expose a method to embed this observation into the surface refinement framework. It should be noticed that this refinement is qualitative, and the geometry that is recovered is not metrically correct. However, augmenting the macroscopic geometry with fine scale features does produce a significant improvement in the perceived quality of the reconstructed face geometry.

For the mesoscopic augmentation, only features that are too small to be recovered by the stereo algorithm are interesting. Therefore, first high pass filtered values are computed for all points X using the projection of a Gaussian $\mathcal{N}$:

$$\mu(\mathrm{X}) = \frac{\sum_{c\epsilon\nu} \alpha_c \left(\mathcal{I}_c(\mathrm{X}) - \left[\mathcal{N}_{\sum_c} \otimes \mathcal{I}_c\right](\mathrm{X})\right)}{\sum_{c\epsilon\nu} \alpha_c} \tag{1.13}$$

where $\mathcal{V}$ denotes the set of visible cameras, $\sum_c$ the covariance matrix of the projection of the Gaussian $\mathcal{N}$ into camera c, and the weighting term α_c is the cosine of the foreshortening angle observed at camera c. The variance of the Gaussian $\mathcal{N}$ is chosen such that high spatial frequencies are attenuated. It can either be defined directly on the surface using the known maximum size of the features or in dependence of the matching window m. The next steps are based on the assumption that variation in mesoscopic intensity is linked to variation in the geometry. For human skin, this is mostly the case. Spatially bigger skin features tend to be smooth and are thus filtered out. The idea is thus to adapt the local high frequency geometry of the mesh to the mesoscopic field (X). The geometry should locally form a concavity whenever (X) decreases and a convexity when it increases.

1.4 Dynamic 3D Face Reconstruction

The objective now is to create dynamic models that accurately recover the facial shape and acquire the time-varying behavior of a real person's face. Modeling facial dynamics is essential for several applications such as avatar animation, facial action analysis, and recognition. Compared with a static or quasi-static object (or scene), this is more difficult to achieve because of the required fast processing. Besides, it is the main limitation of the techniques described in Section 1.3. In particular, laser-based scanners and time-coded structured light shape capture techniques do not operate effectively on fast-moving scenes because of the time required for scanning the object when moving or deforming. In this section, we present appropriate techniques designed for moving/deforming face acquisition and post-processing pipeline for performance capture or expression transfer.

1.4.1 Multiview Dynamic Reconstruction

Passive facial reconstruction has received particular attention because of its potential applications in facial animation. Recent research effort has focused on passive multi-view stereo (PMVS) for animated face capture sans markers, makeup, active technology, and expensive hardware. A key step toward effective performance capture is to model the structure and motion of the face, which is a highly deformable surface. Furukawa and Ponce (2009) proposed a motion capture approach from video stream that specifically aims at this challenge. Assuming that the instantaneous geometry of the face is represented by a polyhedral mesh with fixed topology, an initial mesh is constructed in the first frame using PMVS software for MVS (Furukawa and Ponce, 2010) and Poisson surface reconstruction software (Kazhdan et al., 2006) for meshing. Then its deformation is captured by tracking its vertices $v_1, \ldots v_n$ over time. The goal of the algorithm is to estimate in each frame f the position v_i^f of each vertex v_i (From now on, v_i^f will be used to denote both the vertex and its position.) Each vertex may or may not be tracked at a given frame, including the first one, allowing the system to handle occlusion, fast motion, and parts of the surface that are not initially visible. The three steps of the tracking algorithm refer to local motion parameters estimation, global surface deformation, and filtering.

First, at each frame, an approximation of a local surface region around each vertex, by its tangent plane, gives the corresponding local 3D rigid motion with six degrees of freedom.

Three parameters encode normal information, while the remaining three contain tangential motion information. Then, on the basis of the estimated local motion parameters, the whole mesh is then deformed by minimizing the sum of the three energy terms.

$$\sum_i \left| v_i^f - \hat{v}_i^f \right|^2 + \eta_1 \left| \left[\zeta_2 \Delta^2 - \zeta_1 \Delta \right] \; v_i^f \right|^2 + \; \eta_2 \; E_r \left(v_i^f \right). \tag{1.14}$$

The first *data* term measures the squared distance between the vertex position v_i^f and the position $\hat{v}_i^f$ estimated by the local estimation process. The second uses the discrete Laplacian operator of a local parameterization of the surface in v_i to enforce smoothness. [The values $\zeta_1 = 0.6$ and $\zeta_2 = 0.4$ are used in all experiments (Furukawa and Ponce, 2009)]. This term is very similar to the Laplacian regularizer used in many other algorithms (Ponce, 2008). The third term is also used for regularization, and it enforces local tangential rigidity with no stretch, shrink, or shear. The total energy is minimized with respect to the 3D positions of all the vertices by a conjugate gradient method. In case of deformable surfaces such as human faces, nonstatic target edge length is computed on the basis of non-rigid tangential deformation from the reference frame to the current one at each vertex. The estimation of the tangential deformation is performed at each frame before starting the motion estimation, and the parameters are fixed within a frame. Thus, the tangential rigidity term $E_r(v_i^f)$ for a vertex v_i^f in the global mesh deformation is given by

$$\sum_{v_j \in N(v_i)} \max \left[0, \; \left(e_{ij}^f - \hat{e}_{ij}^f \right)^2 - \; \tau^2 \right], \tag{1.15}$$

which is the sum of squared differences between the actual edge lengths and those predicted from the reference frame to the current frame. The term τ is used to make the penalty zero when the deviation is small so that this regularization term is enforced only when the data term is unreliable and the error is large. In all our experiments, τ is set to be 0.2 times the average edge length of the mesh at the first frame. Figure 1.8 shows some results of motion capture approach proposed in Furukawa and Ponce (2009).

Finally after surface deformation, the residuals of the data and tangential terms are used to filter out erroneous motion estimates. Concretely, these values are first smoothed, and a smoothed local motion estimate is deemed an outlier if at least one of the two residuals exceeds a given threshold. These three steps are iterated a couple of times to complete tracking in each frame, the local motion estimation step only being applied to vertices whose parameters have not already been estimated or filtered out.

The face capture framework proposed by Bradley et al. (2010) operates without use of markers and consists of three main components: acquisition, multiview reconstruction and geometry, and texture tracking. The acquisition stage uses 14 high definition video cameras arranged in seven binocular stereo pairs. At the multiview reconstruction stage, each pair captures a highly detailed small patch of the face surface under bright ambient light. This stage uses on an iterative binocular stereo method to reconstruct seven surface patches independently that are merged into a single high resolution mesh; the stereo algorithm is guided by face details providing, roughly, 1 million polygons meshes. First, depth maps are created from pairs of

Figure 1.8 The results of motion capture approach, proposed by Furukawa and Ponce (2009), form multiple synchronized video streams based on regularization adapted to nonrigid tangential deformation. From left to right, a sample input image, reconstructed mesh model, estimated notion and a texture mapped model for one frame with interesting structure/motion for each dataset 1, 2, and 3. The right two columns show the results in another interesting frame. Copyright © 2009, IEEE

adjacent rectified viewpoints. Observing that the difference in projection between the views causes distortions of the comparison windows, the most prominent distortions of this kind are compensated by a *scaled-window matching* technique. The resulting depth images are converted to 3D points and fused into a single dense point cloud. Then, a triangular mesh from the initial point cloud is reconstructed through three steps: the original point cloud is downsampled using *hierarchical vertex clustering* (Schaefer and Warren, 2003). Outliers and small-scale high frequency noise are removed on the basis of the Plane Fit Criterion proposed by Weyrich et al. (2004) and a point normal filtering inspired by Amenta and Kil (2004), respectively. A triangle mesh is generated without introducing excessive smoothing using *lower dimensional triangulation* methods Gopi et al. (2000).

At the last stage, in order to consistently track geometry and texture over time, a single reference mesh from the sequence is chosen. A sequence of compatible meshes without holes is explicitly computed. Given the initial per-frame reconstructions G_t, a set of compatible meshes M_t is generated that has the same connectivity as well as explicit vertex correspondence. To create high quality renderings, per-frame texture maps T_t that capture appearance changes, such as wrinkles and sweating of the face, are required. Starting with a single reference mesh M_0, generated by manually cleaning up the first frame G_0, dense optical flow on the video images is computed and used in combination with the initial geometric reconstructions G_t to automatically propagate M_0 through time. At each time step, a high quality 2D face texture T_t

from the video images is computed. Drift caused by inevitable optical flow error is detected in the per-frame texture maps and corrected in the geometry. Also, the mapping is guided by an edge-based mouth-tracking process to account the high speed motion while talking.

Beeler et al. (2011) extend their MVS face acquisition system, discussed in Section 1.3, to facial motion capture. Their solution, as Bradley's solution, requires no makeup; the temporally varying texture can be derived directly from the captured video. The computation is parallelizable so that long sequences can be reconstructed efficiently using a multicore implementation. The high quality results derive from two innovations. The first is a robust tracking algorithm specifically adapted for short sequences that integrates tracking in image space and uses the integrated result to propagate a single reference mesh to each target frame. The second is to address long sequences, and it employs the "anchor frame" concept. The latter is based on the observation that a lengthy facial performance contains many frames similar in appearance. One frame is defined as the reference frame. Other frames similar to the reference frame are marked as anchor frames. Finally, the tracker computes the flow from the reference to each anchor independently with a high level of measurement accuracy. The proposed framework operates in five stages:

1. *Stage 1: Computation of Initial Meshes* – Each frame is processed independently to generate a first estimate of the mesh.
2. *Stage 2: Anchoring* – The reference frame is manually identified. Similar frames to the reference frame are detected automatically and labeled as anchor frames.
3. *Stage 3: Image–Space Tracking* – Image pixels are tracked from the reference frame to anchor frames and then sequentially between non-anchor frames and the nearest anchor frame.
4. *Stage 4: Mesh Propagation* – On the basis tracking results from the previous stage, a reference mesh is propagated to all frames in the sequence.
5. *Stage 5: Mesh Refinement* – The initial propagation from Stage 4 is refined to enforce consistency with the image data.

1.4.2 Photometric Stereo

Photometric stereo is a technique in computer vision for estimating the surface normals of objects by observing that object under different lighting conditions. Estimation of face surface normals can be achieved on the basis of photometric stereo assuming that the face is observed under different lighting conditions. For instance, in three-source photometric stereo, three images of the face are given, taken from the same viewpoint and illuminated by three light sources. These light sources emit usually the same light spectrum from three non-coplanar directions. If an orthographic camera model is assumed, the word coordinate system can be aligned so that the xy plane coincides with the image plane. Z axis corresponds to the viewing direction. Hence, the surface in front of the camera can be defined as the height $Z(x, y)$. Now, assuming that ∇z is the gradient of this function with respect to x and y, the vector locally normal to the surface at (x, y) can be defined as

$$n = \frac{1}{\sqrt{1 + |\nabla Z|^2}} \begin{pmatrix} \nabla Z \\ -1 \end{pmatrix}. \tag{1.16}$$

Also, a $2d$ projection operator can be define $P[x] = (x1/x3,\ x2/x3)$ so that it follows that $\nabla z = P[\mathbf{n}]$. The pixel intensity $c_i\,(x,\,y)$ in the ith image, for $i = 1, \ldots 3$, can be defined as

$$c_i\,(x,\,y) = \left(\mathbf{l}_i^T\ \mathbf{n}\right) \int E\,(\lambda) R\,(x,\,y,\,\lambda)\,S\,(\lambda)\,\mathrm{d}\lambda, \tag{1.17}$$

where $\mathbf{l}_i$ is the direction of a light source with spectral distribution $E_i(\lambda)$, illuminating the surface point $(x,\,y,\,z\,(x,\,y))^T$; $R\,(x,\,y,\,\lambda)$ reflectance function, and $S\,(\lambda)$ the response of the sensor camera. The value of this integral is known as Albedo ρ, so the pixel intensity can be defined as

$$c_i = \mathbf{l}_i^{\mathrm{T}}\ \rho\mathbf{n}. \tag{1.18}$$

Using linear constraints of this equation to solve for $\rho\mathbf{n}$ in a least squares sense. The gradient of the height function $\nabla z = P[\rho\mathbf{n}]$ is obtained and integrated to produce the function z. According to three source photometric stereo, when the point is not in shadow with respect to all three lights, three positive intensities c_i can be estimated each of which gives a constraint on $\rho\mathbf{n}$. Thus the following system can be defined as

$$\rho\mathbf{n}{=}\mathbf{L}^{-1}\mathrm{c}. \tag{1.19}$$

If the point is in shadow, for instance in the 1^{st} image, then the estimated of c_1 cannot be used as constraint. In this case, each equation describes a 3D plane, the intersection of the two remaining constraints is a 3D line given by

$$(c_3\,\mathbf{l}_2 - c_2\,\mathbf{l}_3)^{\mathrm{T}}\mathbf{n} = 0. \tag{1.20}$$

In a general case, if the point is in shadow in the ith image, this equation can be arranged as

$$[\mathbf{c}]\,{}^{i}_{\times}\ \mathbf{L}\,\mathbf{n} = \mathbf{0} \tag{1.21}$$

This equation is derived by Wolff and Angelopoulou (1994) and used for stereo matching in a two view photometric. Fan and Wolff (1997) also used this formulation to perform uncalibrated photometric stereo. Hernandez et al. (2011) used that for the first time in a least squares framework to perform three source photometric stereo in the presence of shadows. Figures 1.9 and 1.10 illustrate some reconstruction results with their proposed shading and shape regularization schemes.

1.4.3 Structured Light

Structured light–based techniques are reputed to be precise and rapid. However, 3D imaging of moving objects as faces is a challenging task and usually need more sophisticated tools in combination with the existing patterns projection principle. The first strategy consists in patterns projecting and capturing with a synchronized projecting device and camera at a very high rate. The second is to motion modeling and compensation. Finally, the third fuses several 3D models from one or more projector-camera couples to complete them and corrects sensor

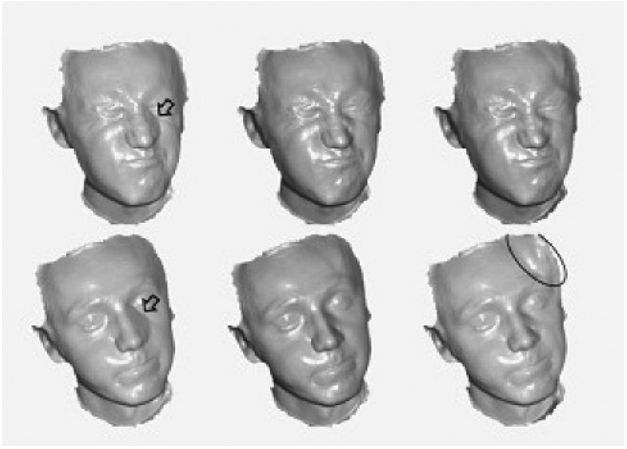

Figure 1.9 Two different frames out of a 1000-frame face video sequence Hernandez et al. (2011). The left column shows the reconstruction when shadows are ignored. Middle and right columns show the corresponding reconstructions after detecting and compensating for the shadow regions using the shading regularization scheme (middle) and shape regularization scheme (right). Note the improvement in the regions around the nose reconstruction where strong cast shadows appear (see arrows). Note also how the shape regularization scheme fails to reconstruct some boundary regions (see circle). Copyright © 2011, IEEE

Figure 1.10 Face sequence. Acquisition of 3D facial expressions based on Hernandez et al. (2007) and the shadow processing technique described in Hernandez et al. (2011). The shadows are processed with the shading regularization scheme. The full video sequence has more than a 1000 frames reconstructed. Copyright © 2011, Springer

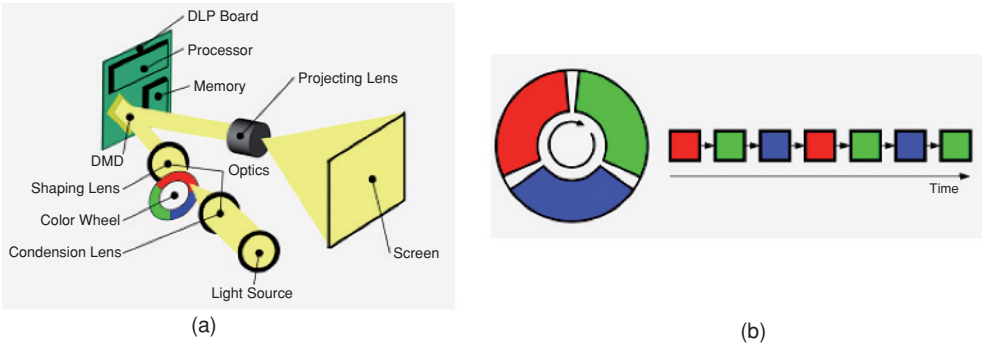

Figure 1.11 (a) DLP projecting technology. (b) Single-chip DLP projection mechanism

errors. These three strategies are presented in the following sections. Pan et al. (2004) have extensively studied the use of color pattern(s) (RGB) and 3-CCD camera. According to their technique, one single color pattern is used, and the data acquisition is fast. If binary or gray-level patterns are used, they must be switched and projected rapidly so that they are captured in a short period. Rusinkiewicz et al. proposed to switch patterns by software (Rusinkiewicz and Levoy, 2001; Rusinkiewicz et al., 2002). To reach fast image switching, Zhang and Yau (2007) proposed to take advantage of the projection mechanism of the single-chip digital-light-processing (DLP) technology. According to their approach, three primary color channels are projected sequentially and repeatedly. This allows capture of three color channel images separately using a synchronized DLP projector device with a digital camera.

A color wheel is a circular disk that spins rapidly. It is composed of R, G, and B filters that color the white light once it passes from in front. Color lights are thus generated. The digital micro-mirror synchronized with the color light, reflects it, and produces three R, G, and B color channel images. Human perception cannot differentiate individual channels as a result of the projection speed. Instead color images are seen. Three phase-shifted sinusoidal patterns are encoded as three primary color channels, R, G, and B of a color image. Three patterns are sent to the single-chip DLP projector from which color filters are removed. A CCD camera is synchronized with the projector and captures each of the three color channels separately into a computer. Unwrapping and phase-to-depth processing steps are applied to the sequence of captured images to recover the depth information. Despite this high speed acquisition, fast motion may still distort the reconstructed phase and hence the reconstructed 3D geometry. Weise et al. (2007) proposed to estimate the error in phase shifting, which produces ripples on the 3D reconstructed surface, and to compensate it. Also, this estimation can provide the motion of the reconstructed 3D surface. Three-step phase shifting has been introduced in Section 1.3 where a sinusoidal pattern is shifted by $\frac{2\pi}{3}$ to produce three patterns, the minimum required to recover depth information:

$$\begin{aligned} I_1(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\cos(\phi(x, y) - \theta), \\ I_2(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\cos(\phi(x, y)), \text{ and} \\ I_3(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\cos(\phi(x, y) + \theta). \end{aligned} \tag{1.22}$$

$I_j,\ j = 1, \ldots 3$, are the recorded intensities, I_0 is the background and I_{mod} is the signal amplitude. $\phi(x, y)$ is the recorded phase value, and θ is the constant phase shift. The phase value corresponds to projector coordinates computed as $\phi = \frac{x_p}{\omega} 2\pi N$, where x_p is the projector x-coordinate, ω the horizontal resolution of the projection pattern, and N the number of periods of the sinusoidal pattern. The wrapped phase is estimated as

$$\hat{\phi}(x, y) = \arctan\left\{\tan\left(\frac{\theta}{2}\right) \frac{I_1(x, y) - I_3(x, y)}{2\, I_2(x, y) - I_1(x, y) - I_3(x, y)}\right\}, \tag{1.23}$$

$$I_0(x, y) = \frac{I_1(x, y) + I_2(x, y) + I_3(x, y)}{3}, \text{ and} \tag{1.24}$$

$$I_{\text{mod}}(x, y) = \sqrt{\frac{(I_3(x, y) - I_0(x, y))^2}{3} + \frac{\left(2I_2(x, y) - I_1(x, y) - I_3(x, y)\right)^2}{9}}. \tag{1.25}$$

Using the estimated phase, the depth is calculated on the basis of triangulation between camera and projection device.

- Motion estimation: Figure 1.12 shows a planar surface and its effects on phase estimation. P_0 is the location observed by the camera at time t_0 and P_1 at t_1. Assuming that $\Delta t = t_0 - t_{-1} = t_1 - t_0$, is a known constant value. If P_0 and P_{-1} are known, the distance vector

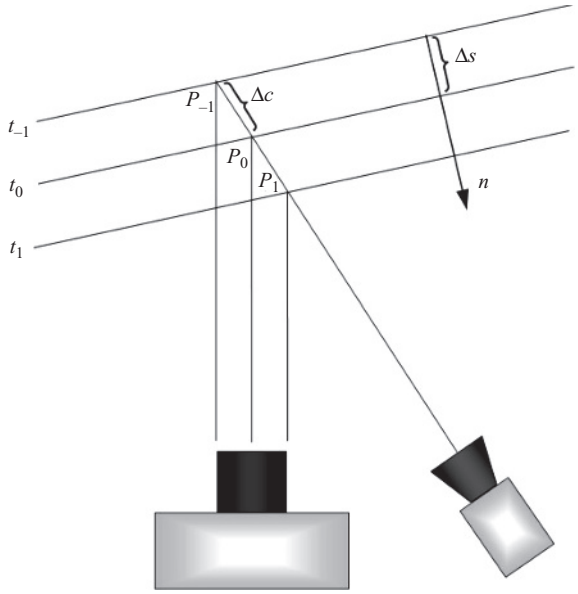

Figure 1.12 A planar surface moving towards the camera and its effect on phase estimation (Weise et al. (2007)). Here three images are captured at three time steps. The velocity of the surface along its normal is estimated on the basis of the normal motion displacement δ_s as the projection of δ_c, the distance vector, onto the surface normal n. Copyright © 2007, IEEE

Δc can be estimated, and thus, the normal motion displacement Δs as the projection of Δc onto the surface normal n. From that, the velocity $\frac{\Delta s}{\Delta t}$ of the surface along its normal can be estimated.

- Error estimation and compensation: Now assume p_0, p_{-1}, and p_1 are projector pixel coordinates of P_0, P_{-1}, and P_1. As the camera and projector are mounted horizontally, the projection pattern is invariant vertically, and only the x-coordinates are of importance. Hence, the difference between the points in the projection pattern is $\Delta x = p_{-1}^x - p_0^x \approx p_0^x - p_1^x$.

As shown earlier, the intensity of an observed pixel in each of the three images depends on I_0, amplitude I_{mod}, phase $\phi(x, y)$, and shift θ. In case of a planar surface, uniform, and diffuse, I_0 and I_{mod} are locally constant on the observed surface. The shift θ is constant. However, as the observed surface is moving, the $\phi(x, y)$ changes between the three images at three different moments in time. At time t_{-1}, t_0, and t_1 camera observes the intensity as projected by p_{-1}, p_0, and p_1, respectively. By converting Δx into the phase difference we have $\Delta\theta = 2\pi N \frac{\Delta x}{\omega}$; Δx being the width of the projection pattern and N the number of projected wrapped phrase. The true intensities are given by

$$\begin{aligned} I_1(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\cos(\phi(x, y) - \theta + \Delta\theta), \\ I_2(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\cos(\phi(x, y)), \text{ and} \\ I_3(x, y) &= I_0(x, y) + I_{\text{mod}}(x, y)\cos(\phi(x, y) + \theta - \Delta\theta). \end{aligned} \tag{1.26}$$

The corrupted shift phase is $\theta - \Delta\theta$. The relative phase error $\Delta\phi$ between observed distorted phase ϕ_d and true phase ϕ_t is

$$\phi_d = \arctan\left(\tan\left(\frac{\theta - \Delta\theta}{2}\right) g\right), \tag{1.27}$$

$$\phi_t = \arctan\left(\tan\left(\frac{\theta}{2}\right) g\right), \tag{1.28}$$

$$\Delta\phi = \phi_d - \phi_t, \text{ and} \tag{1.29}$$

$$g = \frac{I_1(x, y) - I_3(x, y)}{2\, I_2(x, y) - I_1(x, y) - I_3(x, y)}. \tag{1.30}$$

ϕ_t can be expressed as Taylor expansion of ϕ_d:

$$\phi_t = \phi_d + \sin(2\,\phi_d)\; y - \left(\frac{1}{2}\sin(2\,\phi_d) - \frac{1}{4}\sin(4\,\phi_d)\right) y^2 + O\left(y^3\right), \tag{1.31}$$

where $y = \frac{1}{2}\left(\frac{\tan\left(\frac{\theta - \Delta\theta}{2}\right)}{\tan\left(\frac{\theta}{2}\right)} - 1\right)$, $\Delta\theta = \theta - 2\ \arctan\left(\tan\left(\frac{\theta}{2}\right)(2y + 1)\right)$. For small motion, only the first-term of the Taylor expansion is enough. In this case, the undistorted phase values can be locally approximated to evolve linearly along a scanline of the camera: $\phi_t(m) = \phi_c + \phi_m\, m$, where m is the x-coordinate of the pixel. Then a linear least-square

fit can be performed in this local neighborhood (7 pixels used in the author's experiments) of each pixel solving for ϕ_c, ϕ_m, and y:

$$\min_{\phi_c,\ \phi_m,\ \mathrm{y}} \sum \left(\phi_c\left(m\right) - \left(m\phi_m\left(m\right) - \sin\left(2\varphi_d\left(m\right)\right) y\right)\right)^2. \tag{1.32}$$

For large motion, the first-order Taylor degrades, and instead of using the second-order approximation, a faster solution is to use a simulation that estimates y for different values of $\Delta\theta$ and to create a lookup-table (LUT), which is then used to retrieve the true $\Delta\theta$ from an estimated biased y. In this case, a median filter is first applied for robustness. Despite high speed acquisition and motion compensation, imperfections essentially due to sensor noise, residual uncompensated motion and acquisition conditions as illumination may persist. To deal with these problems, Ouji et al. (2011) proposed to apply a 3D temporal super-resolution for each couple of successive 3D point sets $\mathcal{M}_{t-1}$ and $\mathcal{M}_t$ at time t. First, a 3D nonrigid registration is performed. The registration can be modeled as a maximum-likelihood estimation problem because the deformation between two successive 3D faces is nonrigid in general. The coherent point drift (CPD) algorithm, proposed in Andriy Myronenko (2006), is used for the registration the of 3D points set $\mathcal{M}_{t-1}$ with the 3D points set $\mathcal{M}_t$. The CPD algorithm considers the alignment of two point sets $\mathcal{M}_{\mathrm{src}}$ and $\mathcal{M}_{\mathrm{dst}}$ as a probability density estimation problem and fits the Gaussian Mixture Model (GMM) centroids representing $\mathcal{M}_{\mathrm{src}}$ to the data points of $\mathcal{M}_{\mathrm{dst}}$ by maximizing the likelihood as described in Andriy Myronenko (2006). N_{src} is the number of points of m_{src} and $\mathcal{M}_{\mathrm{src}} = \{sn|n = 1, \ldots, N_{\mathrm{src}}\}$. N_{dst} constitutes the number of points of $\mathcal{M}_{\mathrm{dst}}$ and $\mathcal{M}_{\mathrm{dst}} = \{dn|n = 1, \ldots, N_{\mathrm{dst}}\}$. To create the GMM for $\mathcal{M}_{\mathrm{src}}$, a multivariate Gaussian is centered on each point in $\mathcal{M}_{\mathrm{src}}$. All gaussians share the same isotropic covariance matrix $\sigma^2 I$, I being a 3×3 identity matrix and σ^2 the variance in all directions Andriy Myronenko (2006). Hence the whole point set $\mathcal{M}_{\mathrm{src}}$ can be considered as a GMM with the density $p(d)$ as defined by

$$p\left(d\right) = \sum_{m=1}^{N_{\mathrm{dst}}} \frac{1}{N_{\mathrm{dst}}} p\left(d \mid m\right), \quad d \mid m \propto \mathcal{N}\left(\mathrm{s}_m, \sigma^2 \mathrm{I}\right) \tag{1.33}$$

The core of the CPD method is forcing GMM centroids to move coherently as a group, which preserves the topological structure of the point sets as described in Andriy Myronenko (2006). The coherence constraint is imposed by explicit re-parameterization of GMM centroids' locations for rigid and affine transformations. For smooth nonrigid transformations such as expression variation, the algorithm imposes the coherence constraint by regularization of the displacement field Myronenko and Song (2010). Once registered, the 3D points sets $\mathcal{M}_{t-1}$ and $\mathcal{M}_t$ and also their corresponding 2D texture images are used as a low resolution data to create a high resolution 3D point set and its corresponding texture. 2D super-resolution technique as proposed in Farsiu et al. (2004) is applied, which solves an optimization problem of the form:

$$\text{minimize } E_{\mathrm{data}}(H) + E_{\mathrm{regular}}(H). \tag{1.34}$$

The first term $E_{\mathrm{data}}(H)$ measures agreement of the reconstruction H with the aligned low resolution data. $E_{\mathrm{regular}}(H)$ is a regularization or prior energy term that guides the optimizer

Figure 1.13 Some 3D frames computed by the temporal 3D super resolution approach proposed by Ouji et al. (2011)

towards plausible reconstruction H. The 3D model M_t cannot be represented by only one 2D disparity image since the points situated on the fringe change-over have sub-pixel precision. Also, pixels participate separately in the 3D model since the 3D coordinates of each pixel is retrieved using only its phase information. Thus, for each camera three 2D maps are created, defined by the x-, y- and z-coordinates of the 3D points. The optimization algorithm and the deblurring are applied to compute high resolution images of x, y, and z and texture from the low resolution images. The final high resolution 3D point cloud is retrieved by merging obtained 3D models that are already registered since all of them contain the 3D sparse point cloud. The final result is illustrated in Figure 1.13.

1.4.4 Spacetime Faces

The vast majority of stereo research has focused on the problem of establishing spatial correspondences between pixels in a single pair of images for a static moment in time. The works presented in Davis et al. (2003) and Zhang et al. (2003), which presented nearly identical ideas, proposed to introduce the temporal axis (available since they process video sequences) to improve the stereo matching problem. They proposed *spacetime* stereo matching algorithms based on similar ideas. The algorithm proposed in Davis et al. (2003) was tested on

static objects when varying illuminations. The algorithm proposed in Zhang et al. (2003) was tested on moving objects (faces when conveying arbitrary expressions). The following synthesis is based on both works, but the reconstruction results are taken from Zhang et al. (2004) because the object of interest in this chapter is human face. We note that in their experiments, Zhang et al. (2004) used four cameras and two projectors. Each side of the face was acquired by one binocular active stereo system (one projector associated to two cameras). By this way, the authors tried to avoid self-occlusions, which can be a challenging problem in stereo vision (even if a textured light were projected).

- *Spatial stereo matching.* The way in which traditional stereo systems determine the position in space of P, is triangulation, that is by intersection the rays defined by the centers c_l, c_r of cameras C_l, C_r and the projection of P in left and right images $I_l(x_l, y_l, t)$ and $I_r(x_r, y_r, t)$, respectively. Thus triangulation accuracy depends crucially on the solution of corresponding problem. This kind of approaches, widely used in literature, operates entirely within the spatial domain (the images). In fact, knowing the cameras positions ((R, t), the stereo extrinsic parameters), one can first apply rectification transformation that projects left image $I_l(x_l, y, t)$ and right image $I_r(x_r, y, t)$ onto a common image plane, where $y_l = y_r = y$. Thus, the establishing correspondence moves from a 2D search problem to a 1D search problem and minimizes the matching 1D function $F(x_r)$ 1.35, to find x_r^*,

$$F(x_r) = \sum_{V_s} (I_l(V_s(x_l)) - I_r(V_s(x_r)))^2, \tag{1.35}$$

where V_s is a window of pixels in a spatial neighborhood close to x_l (or x_r). The size of V_s is a parameter, it is well-known that the smoothness/noisy reconstruction depends on larger/smaller used window V_s. $F(x_r)$ given in Equation 1.35 is simply the square difference metric. Other metrics exist is the literature, we refer the reader to the review presented in Scharstein and Szeliski (2002). Figure 1.15**c** shows the reconstructed facial surface from *passive stereo* (left top frame is given Fig. 1.15**a**). Here, neither light pattern is projected on the face. The reconstruction result is noisy due to the texture homogeneity on the skin regions, which leads to matching ambiguities. In contract, an improved reconstruction is

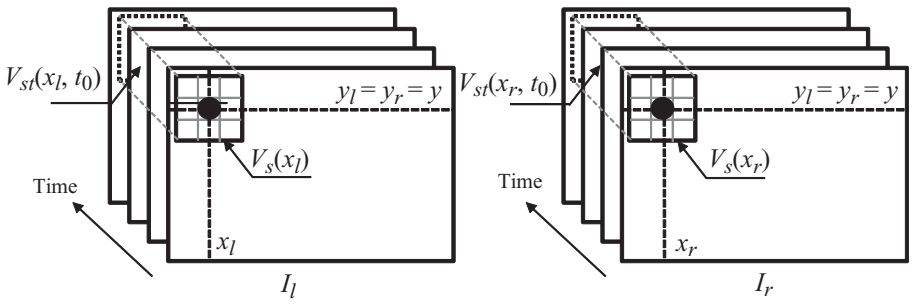

Figure 1.14 Spatial vs. Spacetime stereo matching. The spatial matching uses only spatial axis along y, thus the V_s window to establish correspondence. The spacetime stereo matching extend the spatial window to the time axis, thus the V_{st} is used to compute $F(x_r)$

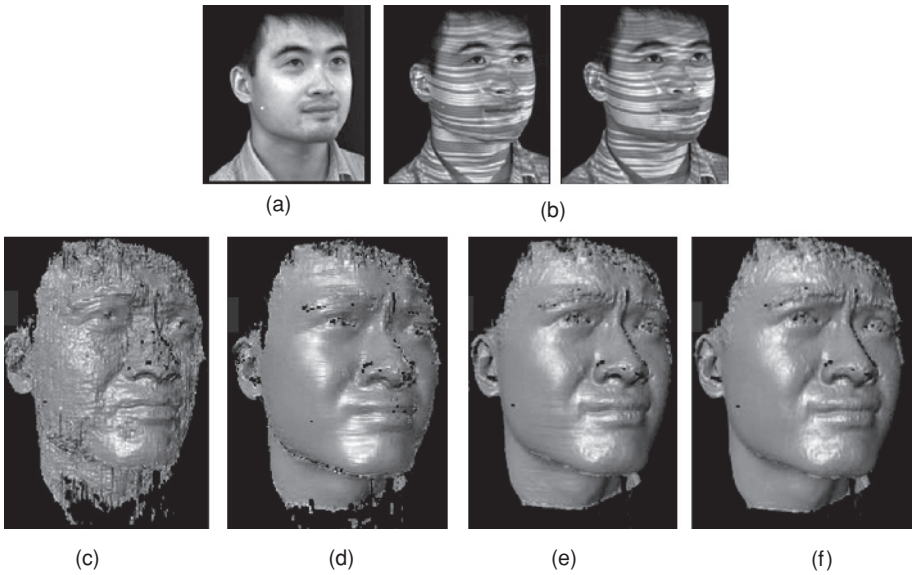

Figure 1.15 Comparison of four different stereo matching algorithms. (a) Top left non-pattern frame, captured in ambient lighting conditions. (b) Sequence of top left pattern frames, captured under patterns projections. (c) Reconstructed face using traditional stereo matching with a [15 × 15] window achieved on non-pattern left stereo frames. The result is noisy due to the lack of color variation on the face. (d) Reconstructed face using pattern frames (examples are given in (b)) using stereo matching with a [15 × 15] window. The result is much better because the projected stripes provide texture. However, certain face details are smoothed out due to the need for a large spatial window. (e) Reconstructed face using local spacetime stereo matching with a [9 × 5 × 5] window. (f) Reconstructed face using the global spacetime stereo matching with a [9 × 5 × 5] window. Global spacetime stereo matching removes most of the striping artifacts while preserving the shape details [from http://grail.cs.washington.edu/projects/stfaces/]

given in Figure 1.15**d**, where *active stereo* is used. The projected colored stripes generate texture on the face, which helps the spatial matching process. However, certain facial shape details are smoothed out because of the largeness of the used spatial window (15 × 15). Frames shown in Figure 1.15**b**) illustrate pattern projections on the face across time.

- *Temporal stereo matching.* In this stereo-matching schema, establishing correspondence for a pixel (x_l, y, t_0) in frame $\mathcal{M}$ is based, this time, on temporal neighborhood $V_t = t_0 \pm \Delta t$, instead of the spatial window V_s. Thus, one can define the matching function $F(x_r)$ as follows,

$$F(x_r) = \sum_{V_t} (I_l(V_t(x_l, t_0)) - I_r(V_t(x_r, t_0)))^2 \tag{1.36}$$

The previous equation is analogous to Equation 1.35 except that now instead of a spatial neighborhood, one must consider a temporal neighborhood V_t around some central time t_0. Because of the changing of the light patterns over time, this temporal window works. This

time, the size of V_t is a parameter, that is, the accuracy/noisy reconstruction depends on larger/smaller of the used window. It should be also adapted to deforming objects speed.

- *Spacetime stereo matching.* This stereo-matching schema combines both spatial matching and temporal one to limit the matching ambiguities. The function $F(x_r)$ is analogous to Equations 1.35 and 1.36 and is given by

$$F(x_r) = \sum_{V_{st}} (I_l(V_{st}(x_l, t_0)) - I_r(V_{st}(x_r, t_0)))^2, \tag{1.37}$$

Here V_{st} represents a spatiotemporal volume instead of a window in a spatial-based matching or a vector in a temporal-based matching. Figure 1.14 illustrates the spatial and the spacetime stereo matchings to establish correspondences between the pixels in I_l and those in I_r. The images are already rectified. Figure 1.15**e** shows the reconstruction result operated by spatio-temporal stereo matching using a volume of size $(9 \times 5 \times 5)$. This time, the spacetime approach cover more shape details than in Fig. 1.15**d**, however, it also yields artifacts due to the over-parametrization of the depth map. An improvement of this reconstruction using a global spacetime stereo matching with the same volume size is given in Fig. 1.15**f**. (See[2] for video illustrations of these reconstructions).

1.4.5 Template-based Post-processing

Recently, template-based approaches emerge due to its simplicity and robustness to noisy range data. Outputs of shape recovery techniques present often imperfections like spikes, holes dues to self-occlusions or the absorption of projected lights by dark regions of the face. The template generic model provides a strong geometric prior and thus leads to high quality reconstructions with automated hole-filling and noise removal. Correspondence estimation is often facilitated by the use of tracked marker points or hand-selected landmarks correspondences. The template-based literature consist on template-to-data registration then fitting and could allowing 3D face tracking and expressions cloning. These stages are described in detail in the following paragraphs. For the rest of this section, let $\mathcal{M}$ denotes the template model and $\mathcal{P}$ denotes the target data.

Landmarks Detection

This step consists on manually or automatically facial *keypoints* detection (eyebrows, eyes, nose, mouth contours, etc.). These facial *keypoints* are important in the following stages. In particular, they could be used in coarse rigid registration to prepare the fine one, and they are often used as control points in the warping/fitting procedure. Automatic 3D face landmarking is one active research topic studied within the 3D face recognition and expression recognition applications. Many approaches are designed and try to face the pose variation and external occlusion problems (Segundo et al., 2010; Mehryar et al., 2010; Zhao et al., 2011).

[2] http://grail.cs.washington.edu/projects/stfaces/

Rigid Registration

Registration of $\mathcal{M}$ and $\mathcal{P}$ involves estimating an optimal rigid transformation between them, denoted T. Here, $\mathcal{P}$ is assumed to remain stationary (the reference data), whereas $\mathcal{M}$ (the source data) is spatially transformed to match it. The *Iterative Closet Point* algorithm (ICP) is the best-known technique for pairwise surface registration. Since the first paper of Besl and McKay (1992) ICP has been widely used for geometric alignment of 3D models and many variants of ICP have been proposed (Rusinkiewicz and Levoy, 2001). ICP is an iterative procedure minimizing the error (deviation) between points in $\mathcal{P}$ and the closest points in $\mathcal{M}$. It is based one of the following two metrics: (i) *the point-to-point*, metric which is the earlier and the classical one, by minimizing in the k-th iteration, the error $E^k_{\text{reg}}(T^k) = \sum(T^k.p_i - q_j)$; $q_j = q|\min_{q\in\mathcal{M}}(E^k_{\text{reg}}(T^k))$; (ii) *the point-to-plane* introduced later and minimizes $E^k_{\text{reg}}(T^k) = \sum n(q_j)(T^k.p_i - q_j)$. For each used metrics, this ICP procedure is alternated and iterated until convergence (i.e., stability of the error). Indeed, total transformation T is updated in an incremental way as follows: for each iteration k of the algorithm: $T = T^k.T$. One note that ICP performs fine geometric registration assuming that a coarse registration transformation T^0 is known. The final result depends on the initial registration. The initial registration could be obtained when corresponding detected landmarks in $\mathcal{M}$ and $\mathcal{P}$.

Template Warping/Fitting

A warping of $\mathcal{M}$ to $\mathcal{P}$ is defined as the function F such that $F(\mathcal{M}) = \mathcal{P}$. The function F is called the *warping function*, which takes $\mathcal{M}$ to $\mathcal{P}$. Given a pair of landmarks (detected as described in Section 1.4.5) with known correspondences, $U_L = (u_i)^T_{1<i<L}$ and $V_L = (v_i)^T_{1<i<L}$, in $\mathcal{M}$ and $\mathcal{P}$, respectively. One needs to establish dense correspondence between other meshes vertices; u_k and v_k denote the locations of the k-th corresponding pair and L is the total number of corresponding landmarks. Thus, a warping function, F, that warps U_L to V_L subject to perfect alignment is given by the conditions $F(u_i) = v_i$ for $i = 1, 2, \ldots, L$.

- *Thin Plate Spline (TPS)*. TPS Bookstein (1989) are a class of widely used non-rigid interpolating (warping) functions. The thin plate spline algorithm specifies the mapping of points for a reference, $\mathcal{P}$, set to corresponding points on a source set, $\mathcal{M}$. The TPS fits a mapping function $F(u)$ between corresponding point-sets $\{v_i\} \in \mathcal{M}$ and $\{u_i\} \in \mathcal{P}$ by minimizing the following energy function:

$$E_{tps} = \sum_{i=1}^{L} \|v_i - F(u_i)\|^2 + L\lambda J \tag{1.38}$$

For a fixed λ which provides trade-off of warp smoothness and interpolation.

$$J = \iint \left[\left(\frac{\partial^2 F}{\partial u^2} \right)^2 + 2 \left(\frac{\partial^2 F}{\partial u \partial v} \right)^2 + \left(\frac{\partial^2 F}{\partial v^2} \right)^2 \right] \mathrm{d}u\, \mathrm{d}v \tag{1.39}$$

The interpolation deformation model is given in terms of the warping function $F(u)$, with

$$F(u) = \overbrace{A}^{4\times 4} u + \underbrace{W^T}_{4\times L} \overbrace{K(u)}^{L\times 1}, \tag{1.40}$$

where A (affine transformation) and W (non-affine warping) are TPS parameters and $K(u) = (|u - u_1|; |u - u_2|; \ldots; |u - u_m|)^T$ is the control point influence vector.

The warping coefficients A and W are computed by the equation:

$$(A|W) \left(\begin{array}{c|c} U & 0 \\ \hline K + L\lambda I & U^T \end{array} \right) = (V|0) \tag{1.41}$$

In other words, any point on $\mathcal{M}$ close to a source landmark v_k will be moved to a place close to the corresponding target landmark u_k in $\mathcal{P}$. The points in between are interpolated smoothly using Bookstein's Thin Plate Spline algorithm Bookstein (1989).

- *Non-rigid ICP*. Register in a non-rigid way a template $\mathcal{M}$ and an input scan $\mathcal{P}$ by non-rigid ICP requires estimating both correspondence and a suitable warping function that matches the shape difference between them. In Allen et al. (2003) and Amberg et al. (2007) similar ideas are presented for scan-template warping applied on human body in Allen et al. (2003) and on human faces in Amberg et al. (2007). Both of them proposed an energy-minimization framework, as given by

$$E = \alpha \overbrace{\sum_{v_i \in \mathcal{M}} w_i \text{dist}^2(T^i v_i, \mathcal{P})}^{E_{\text{data}}(T)} + \beta \overbrace{\sum_{i,j|\{vi,vj\}\in \text{edges}(\mathcal{M})\}} \|T^i - T^j\|_F^2}^{E_{\text{smoothness}}} + \gamma \overbrace{\sum_i \|T^i v_i - u^j\|^2}^{E_{\text{landmarks}}}, \tag{1.42}$$

where minimizing the term E_{data} guarantee that the distance between the deformed template $\mathcal{M}$ and the target data $\mathcal{P}$ is small. The term $E_{\text{smoothness}}$ is used to regularize the deformation. In other words, it penalizes large displacement differences between neighboring vertices. The term $E_{\text{landmarks}}$ is introduced to guide the deformation by using corresponding control points that are simply the anthropometric markers in human body and facial landmarks in the case of face fitting. Similar formulation are presented in Zhang et al. (2004) for template fitting. The Figure 1.16 illustrates an example of template fitting results. A similar formulation is used in Weise et al. (2009) for personalized template building.

Template Tracking

In Zhang et al. (2004), after the template fitting step, the authors proposed a tracking procedure which yields point correspondence across the entire sequence. They obtained time-varying face models (of the deformed template) without using markers. Once this template sequence is acquired, they propose to interactively manipulate it to create new expressions. To achieve template tracking, they first compute optical flow from the sequence. The flow represents

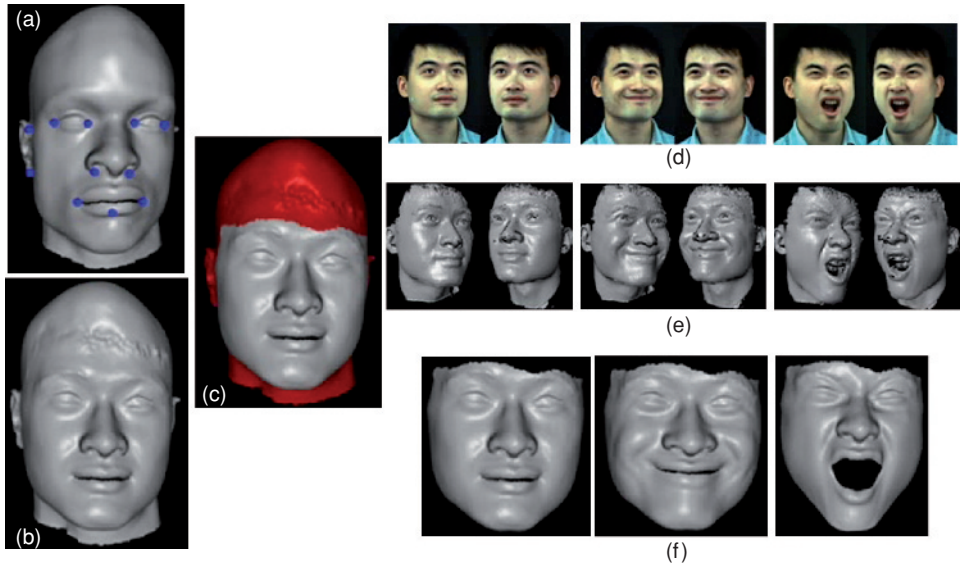

Figure 1.16 Illustration of the template fitting process. (a) A face template with manual landmarks. (b) Obtained mesh after fitting the warped template to the first two depth maps given in (e). (c) Facial region limitation (red colored regions present unreliable depth or optical flow estimation). (d) A sequence of texture image pairs. (e) A sequence of depth map pairs. (f) Selected meshes after tracking the initial mesh through the whole sequence, using both depth maps and optical flows. The process is marker-less and automatic [from http://grail.cs.washington.edu/projects/stfaces/]

vertices motion across the facial sequence and is used to enhance template tracking by establishing inter-frame correspondences with video data. Then, they measure the consistency of the optical flow and the vertex inter-frame motion by minimizing the defined metric. Similar ideas were presented in Weise et al. (2009) where a person-specific facial expression model is constructed from the tracked sequences after non-rigid fitting and tracking. The authors targeted real-time puppetry animation by transferring the conveyed expressions (of an actor) to new persons. In Weise et al. (2011) the authors deal with two challenges of performance-driven facial animation; accurately track the rigid and non-rigid motion of the user's face, and map the extracted tracking parameters to suitable animation controls that drive the virtual character. The approach combines these two problems into a single optimization that estimates the most likely parameters of a user- specific expression model given the observed 2D and 3D data. They derive a suitable probabilistic prior for this optimization from pre-recorded animation sequences that define the space of realistic facial expressions.

In Sun and Yin (2008), the authors propose to adapt and track a generic model to each frame of 3D model sequences for dynamic 3D expression recognition. They establish the vertex flow estimation as follows: First, they establish correspondences between 3D meshes using a set of 83 pre-defined key points. This adaptation process is performed to establish a matching between the generic model (or the source model) and the real face scan (or the target model). Second, once the generic model is adapted to the real face model, it will be considered as an intermediate tracking model for finding vertex correspondences. The vertex

correspondence across 3D model sequences provides a set of motion trajectories (vertex flow) of 3D face scans. The vertex flow can be depicted on the adapted generic model (tracking model) through the estimation of the displacement vector from the tracked points of the current frame to the corresponding points of the first frame with a neutral expression. The vertex flow is described by the facial motion vector $U = [u_1, u_2, \ldots, u_n]$, where n is the number of vertices of the adapted generic model. They used the Hidden Markov Model to model and train facial dynamics.

Expression Transferring

Also known as *expression cloning* or *performance capture* when facial animation uses the performance of an actor to animate virtual models. The steps discussed earlier, namely, template fitting and tracking, allow expression transferring from real-time acquired 3D data to a virtual model or puppetry. Several papers were published to transfer facial animation to templates, puppetry or personalized models, yong Noh and Neumann (2001), Sumner and Popović (2004), Vlasic et al. (2005), Pyun et al. (2003), Zhang et al. (2004), Weise et al. (2009), Weise et al. (2011), etc.

1.5 Summary and Conclusions

Creating 3D face models that look and deform realistically in an important issue is many applications such as person-specific facial animation, 3D-based face recognition, and 3D-based expression recognition. This chapter is a survey of successful state-of-the-art techniques that sometimes led to commercial systems. These techniques are within a static/dynamic (moving) face modeling–guided taxonomy. Each of the presented techniques is based on one of the following well-known concepts: (i) depth from triangulation, (ii) shape from shading, and (iii) depth from ToF. Obviously, other approaches exist in the literature but we limited our survey to those based on the aforementioned concepts. In this section, we will put forward, a comparative study of the mentioned approaches according to the *intrinsic* and *extrinsic* factors. The *intrinsic factors* are related to the sensor, such as its cost, its spatial (in the case of static modeling) or spatio-temporal resolutions (in the case of dynamic modeling), its measurement accuracy, and its intrusiveness/need user cooperation. The *extrinsic factors* include variations due to illumination changes, motion speed of the observed face, and details in the face (wrinkles, scars, etc.). Figure 1.17 gives an evaluation of approaches according to these criteria.

☞ Laser-stripe scanning is intended for *static faces* due to the processing time required to project the laser stripe on the whole face. The sensor is expensive and needs user cooperation to perform face acquisition (a distance less than 1.5 m is required). Commercial systems such as the Minolta Non-contact 3D Digitizer VIVID-910[3] produced texture and depth images of the same resolution 640 × 480. The system accurately measures the 3D object with a depth-accuracy of around 0.1 mm. It takes 2.5 s for the fine mode and 0.5 s for the fast mode to produce a scan, thus no motion during the scan is tolerated. Laser-based techniques

[3]http://www.konicaminolta.com/instruments/download/catalog/3d/pdf/vivid910_e9.pdf

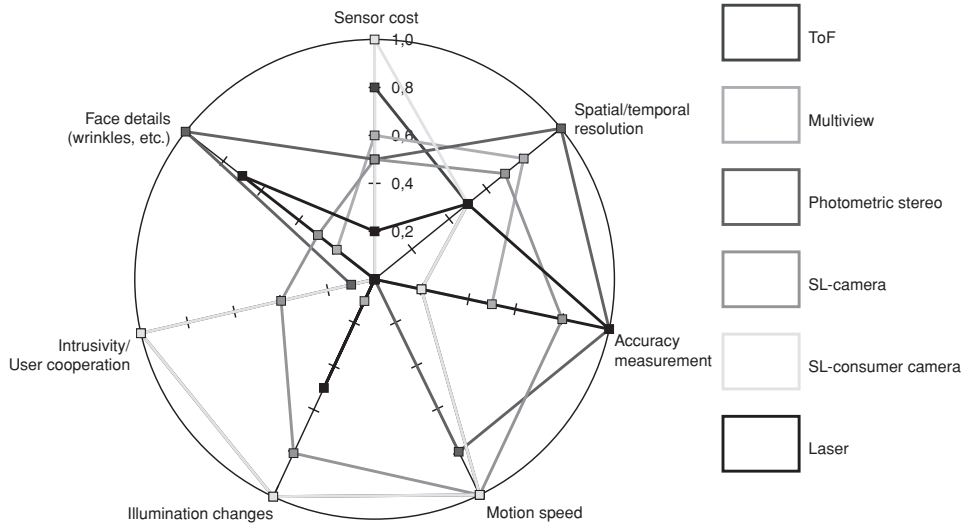

Figure 1.17 Evaluation of different 3D face modeling techniques according to *Extrinsic factors* (Motion speed/illumination changes/intrusivity and need for user cooperation/face details) and Intrinsic factors (spatial and temporal resolutions/accuracy measurement/sensor cost)

work in office environment lighting conditions. Most of the facial details (wrinkles, scars, and other person-specific markers) are reproduced in the virtual model.

☞ Structured-light (SL) techniques provided an attractive alternative to the expensive laser-stripe scanning technique. In fact, projected light(s) intend(s) to replace the laser scan. The ARTEC MHT 3D scanner[4] is one commercial system which projects a permanent light pattern and produces 3D video of a rate around 15 fps. Frame resolution is about 500,000 points. The working distance should be in the interval of 0.4–1 m. Depth-measurement accuracy is comparable to laser scanners and is about 0.1 mm. Texture channel is also captured but only when needed. The sensor is cheaper than the laser digitizers.

☞ Structured-light (SL) consumer depth cameras (which are much cheeper) as MS Kinect[5] and Asus Xtion Pro Live[6] have been recently developed and have been an attractive alternative for expensive sensors. Kinect is based on the permanent projection of one infrared-laser pattern. It was primarily designed for natural interaction in a computer game environment. In fact, the sensor is less intrusive and only a near-frontal position of the user is needed. However, the characteristics of the data captured by Kinect have attracted the attention of researchers in the field of computer vision and computer graphics. The camera provides depth and texture video with 300,000 points in every frame. The 2D and 3D videos have got a rate of 30 fps. The depth measurement produced by the Kinect was not so accurate, which means that it achieves a coarse reconstruction of the 3D face.

[4]http://www.artec3d.com/3d_scanners/artec-mht
[5]http://www.xbox.com/fr-fr/kinect
[6]http://www.asus.com/Multimedia/Motion_Sensor/Xtion_PRO_LIVE/

- ☞ In photometric stereo approaches, only one image of the face is captured and used to recover the depth information. The face is illuminated with three colored light sources from three different directions. The capture can be made in real time enabling 3D and 4D acquisition. By knowing the surface reflectance properties of the face, the local surface orientation at points illuminated by all three light sources are computed. One of the important advantages of the photometric stereo approaches is that the points do not need to be registered. Thus, this category of approaches does not suffer from the correspondence problem, providing high performance for featureless surfaces as the human skin. On the other hand, the disadvantages of this category of approaches are that they are indirect and practical only for applications in which the illumination is carefully controlled.
- ☞ Multiview-based approaches capture images instantly and provide high resolution 3D and 4D textured images. In addition, they have several advantages over active approaches. First is the quality texture image. The acquisition phase does not require pattern projection, and, so, there is a true one-to-one correspondence with every color pixel and every 3D point. The original texture images are always of the highest quality. Second is the absence of holes in the final 3D scans. Dimensional Imaging[7] proposes systems designed specifically to capture high definition 3D surface images of the human face with highly detailed 20-megapixel texture maps using four 10-megapixel cameras and up to 32 cameras with a resolution of up to 21 megapixels.
- ☞ ToF cameras are relatively new devices, as the semiconductor processes have only recently become fast enough for such devices. The systems cover ranges of a few meters up to about 60 m. Another advantage of ToF systems is the high rate capture. In return, they have a low resolution and a precision of 1 mm to 1 cm. The Mesa Imaging[8] SwissRanger 4000 (SR4000) is probably the most well-known ToF camera. It has a range of 5–8 m, 176 × 144 pixel resolution over 43.6° × 34.6° field of view and operates at up to 54 fps.The PMD Technologies[9] CamCube 2.0 is a less popular one. It has a range of 7 m, 204 × 204 pixel resolution with 40.0° × 40.0° field of view. It operates at 25 fps.

Exercises

1. From Figure 1.18 prove that $\frac{AB}{A'B'} = \frac{AC}{A'C'} = \frac{BC}{B'C'} = \frac{h_{ABC}}{h_{A'B'C'}}$; retrieve the Z formula given in Equation 1.8.

2. We need to study the 3D scanning prototype given in Figure 1.19. It consists of a laser source that illuminates the object to be continuously scanned and two cameras that look at the same object. The projected laser stripe is seen by both cameras. The global sensor calculates the depth information, as illustrated in the figure. To capture the full geometry of the object, a manual scan of the surface is required.
- Compute the Z_1-coordinate together with Z_2-coordinate.
- Explain the triangles considered to calculate Z_1-coordinate.
- Why this prototype involves two sensors, each of them capable of measuring the depth. Suggest a depth value Z as a function of Z_1 and Z_2; explain your choice.

[7] http://www.di3d.com
[8] http://www.mesa-imaging.ch/
[9] http://www.pmdtec.com/

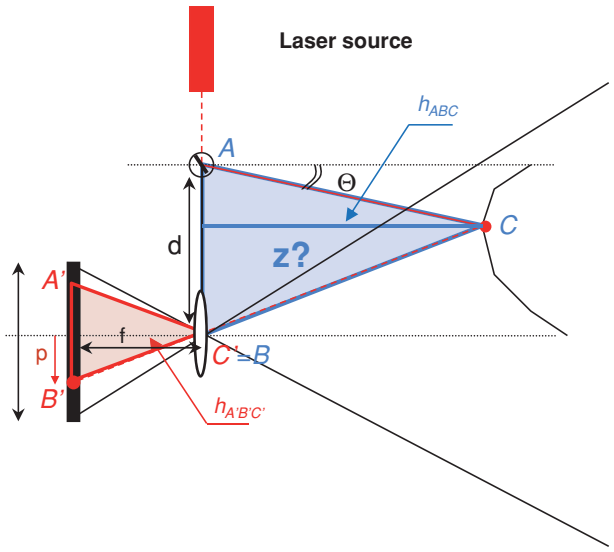

Figure 1.18 Optical triangulation geometry (laser stripe scanning)

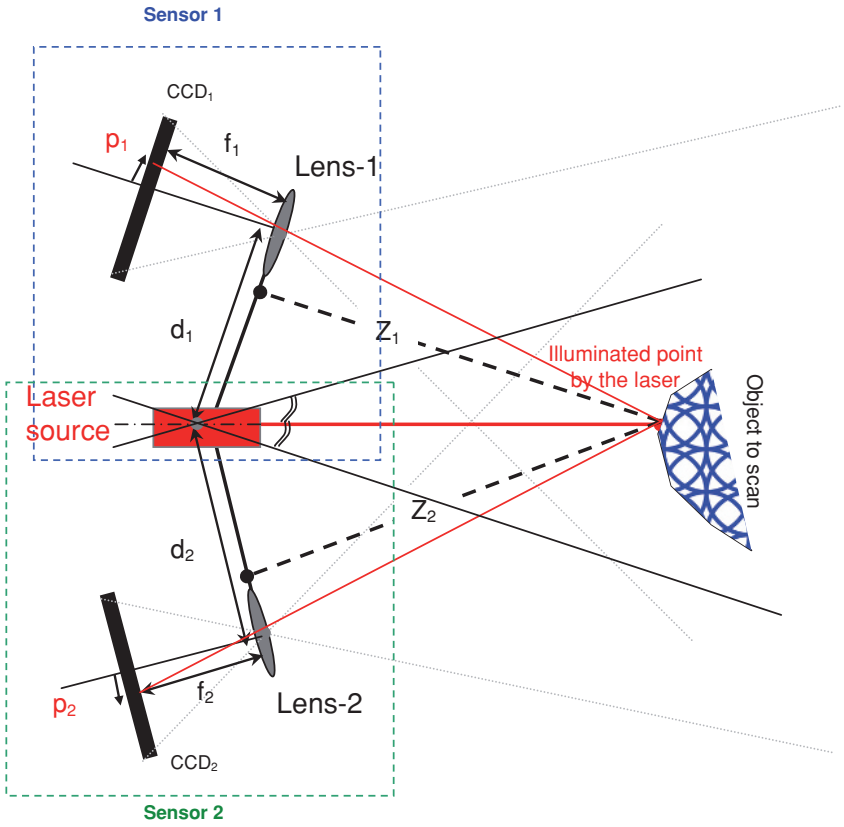

Figure 1.19 3D scanning system to study

3. We aim to produce patterns associated with Equation 1.10. Assuming that the gray level 0 represents black and gray level 255 represents white, find the value of I_0 and I_{mod}. Solve Equation 1.10 for I_0, I_{mod}, and Equation 1.10.

4. Extend Equation 1.10 to obtain N patterns. Solve these equations for I_0, I_{mod}, and Equation 1.10. What is the impact of using more than three patterns on the accuracy and delay?

5. Assuming that only one pattern from Equation 1.10 is used, resolve I_0, I_{mod}, and Equation 1.10 using Fourier Transfrom.

6. The approach presented in Section 1.4, Equations 1.27 to 1.32, overcomes the major problem of fast phase-shift scanning, namely, motion artifacts. An analysis of the motion error has been introduced to compensate for motion artifacts on the pixel level. Nevertheless, high-frequency texture can still pose problems during motion, as the assumption of invariant surface reflectance is violated. Investigate the possibility of adding a stereo module and extending the motion compensation to stereo geometry to handle these cases.

References

Allen B, Curless B, Popović Z. The space of human body shapes: reconstruction and parameterization from range scans. ACM Transactions on Graphics 2003;22(3):587–594.

Amberg B, Romdhani S, Vetter T. Optimal step nonrigid icp algorithms for surface registration. 2007;CVPR. IEEE Computer Society.

Amenta N, Kil YJ. Defining point-set surfaces. ACM SIGGRAPH 2004 Papers on SIGGRAPH 04 2004;23(3):264.

Andriy Myronenko, Xubo B. Song MÁCP. Nonrigid point set registration: coherent point drift. NIPS 2006;1009–1016.

Beeler T, Bickel B, Beardsley P, Sumner B, Gross M. High-quality single-shot capture of facial geometry. ACM Transactions on Graphics 2010;29(4):40:1–40:9.

Beeler T, Hahn F, Bradley D, Bickel B, Beardsley P, Gotsman C, Sumner RW, Gross M. High-quality passive facial performance capture using anchor frames. ACM SIGGRAPH 2011 papers on SIGGRAPH11 2011;1(212):1.

Besl PJ, McKay ND. A method for registration of 3-d shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 1992;14(2):239–256.

Bookstein FL. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 1989;11(6):567–585.

Bradley D, Heidrich W, Popa T, Sheffer A. High resolution passive facial performance capture. ACM Transactions on Graphics 2010;29:41:1–41:10.

Davis J, Ramamoorthi R, Rusinkiewicz S. Spacetime stereo: a unifying framework for depth from triangulation. IEEE Computer Vision and Pattern Recognition 2003;2:359–366.

Fan J, Wolff LB. Surface curvature and shape reconstruction from unknown multiple illumination and integrability. Journal of Computer Vision and Image Understanding 1997;65(2):347–359.

Farsiu S, Robinson MD, Elad M, Milanfar P. Fast and robust multiframe super resolution. IEEE Transactions on Image Processing 2004;13(10):1327–1344.

Furukawa Y, Ponce J. Accurate, dense, and robust multi-view stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence 2010;32(8):1362–1376. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5226635&tag=1.

Furukawa Y, Ponce J. Dense 3D motion capture for human faces. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL: Computer Vision and Pattern Recognition; 2009. p. 1674–1681. Available at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5206868.

Gargallo P, Sturm P. Bayesian 3d modeling from images using multiple depth maps. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2005 Jun 20–26; San Diego, CA: Computer Vision and Pattern Recognition; 2005. 2(c). p. 885–891.

Glencross M, Ward GJ, Melendez F, Jay C, Liu J, Hubbold R. A perceptually validated model for surface depth hallucination. ACM Transactions on Graphics 2008;27(3):1.

Gopi M, Krishnan S, Silva CT. Surface reconstruction based on lower dimensional localized Delaunay triangulation. Computer Graphics Forum 2000;19(3):467–478.

Hernandez C, Vogiatzis G, Cipolla R. Overcoming shadows in 3-source photometric stereo. Pattern Analysis and Machine Intelligence, IEEE Transactions on 2011;33(2):419–426.

Hernandez C, Vogiatzis G, Brostow G, Stenger B, Cipolla R. Non-rigid photometric stereo with colored lights. IEEE 11th International Conference on Computer Vision (2007) 2007;0(5):1–8.

Hernandez Esteban C, Schmitt F. Silhouette and stereo fusion for 3d object modeling. Computer Vision and Image Understanding 2004;96(3):367–392.

Kazhdan M, Bolitho M, Hoppe H. Poisson surface reconstruction. Proceedings of the 4th Eurographics Symposium on Geometry Processing; 2006 Jun 26–28; Cagliari, Sardinia, Italy: Eurographics Symposium on Geometry Processing; 2006. p. 61–70.

Kolmogorov V, Zabih R. Multi-camera scene reconstruction via graph cuts. European Conference on Computer Vision 2002;8:82–96.

Mehryar S, Martin K, Plataniotis KN. Automatic landmark detection for 3d face image processing. IEEE Congress on Evolutionary Computation 2010;1–7.

Morris DD, Kanade T. Image-consistent surface triangulation. Robotics 2000;1:332–338.

Myronenko A, Song X. Point set registration: coherent point drift. IEEE Transactions on Pattern Analysis and Machine Intelligence 2010;32(12):2262–2275.

Narayanan PJ, Rander PW, Kanade T. constructing virtual worlds using dense stereo. Proceedings of the 6th International Conference on Computer Vision; 1998 Jan 4–7; Mumbai, India: Narosa Publishing House; 1998. p. 3–10.

Ouji K, Ardabilian M, Chen L, Ghorbel F. Multi-camera 3d scanning with a non-rigid and space-time depth super-resolution capability. In: Pedro Real, Daniel Diaz-Pernil, Helena Molina-Abri, et al., editors. Computer Analysis of Images and Patterns. Proceedings of the 14th international conference on computer analysis of images and patterns, Part II; 2011 Aug 29–31 Seville, Spain. Berlin, Heidelberg: Springer-Verlag; 2011. 220–228.

Pan J, Huang P, Zhang S, Chiang FP. Color n-ary gray code for 3-d shape measurement. Proceedings of the 12th International Conference on Experimental Mechanic; 2004 Aug 29–Sep; Bari, Italy.

Ponce J. Dense 3d motion capture from synchronized video streams. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 5(Image and Geometry Processing for 3-D Cinematography); 2008 Jun 24–26; Anchorage, AK: IEEE; 2008. p. 1–8.

Pyun H, Kim Y, Chae W, Kang HW, Shin SY. An example-based approach for facial expression cloning. Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on computer animation, SCA'03; Aire-la-Ville, Switzerland: Eurographics Association; 2003. p. 167–176.

Rusinkiewicz S, Levoy M. Efficient variants of the ICP algorithm. Proceedings of 3rd International Conference on 3D Digital Imaging and Modeling (3DIM); 2001 May 29–Jun 1; Quebec City, Canada: IEEE Computer Society; 2001. p. 145–152.

Rusinkiewicz S, Hall-Holt O, Levoy M. Real-time 3d model acquisition. ACM Transactions on Graphics 2002;21(3):438–446.

Schaefer S, Warren J. Adaptive vertex clustering using octrees. Proceedings of SIAM Geometric Design and Computation; 2003; New York: SIAM. 2003; p. 491–500.

Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 2002;47(1–3):7–42.

Segundo MP, Silva L, Bellon ORP, Queirolo CaC. Automatic face segmentation and facial landmark detection in range images. IEEE Transactions on Systems, Man, and Cybernetics Part B 2010;40(5):1319–1330.

Seitz SM, Dyer CR. Photorealistic scene reconstruction by voxel coloring. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1997;35(2):1067–1073.

Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R. A comparison and evaluation of multi-view stereo reconstruction algorithms. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Volume 1 CVPR06 2006;1(c):519–528.

Sumner RW, Popović J. Deformation transfer for triangle meshes. ACM Transactions on Graphics 2004;23(3):399–405.

Sun Y, Yin L. Facial expression recognition based on 3d dynamic range model sequences. Proceedings of the 10th European Conference on Computer Vision: Part II. Berlin, Heidelberg: Springer-Verlag: ECCV '08; 2008. p. 58–71.

Taylor CJ. Surface Reconstruction from feature based stereo. Proceedings of the 9th IEEE International Conference on Computer Vision; 2003. p. 184–190.

Treuille A, Hertzmann A, Seitz SM. Example-based stereo with general BRDFs Volume 2; 2004; Berlin: Springer-Verlag; 2004. p. 457–469.

Vlasic D, Brand M, Pfister H, Popovic J. Face transfer with multilinear models. ACM Transactions on Graphics 2005;24(3):426–433.

Weise T, Bouaziz S, Li H, Pauly M. Realtime performance-based facial animation. ACM Transactions on Graphics (Proceedings SIGGRAPH 2011); 2011.

Weise T, Leibe B, Van Gool L. Fast 3d scanning with automatic motion compensation. IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07; 2007. p. 1–8.

Weise T, Li H, Gool LV, Pauly M. Face/off: Live facial puppetry. Proceedings of the 2009 ACM SIGGRAPH/ Eurographics Symposium on Computer animation (Proc. SCA'09). Zurich:Eurographics Association, ETH; 2009.

Weyrich T, Pauly M, Heinzle S, Keiser R, Scandella S, Gross M. Post-processing of scanned 3d surface data. Science 2004;1:85–94.

Wolff LB, Angelopoulou E. Three-dimensional stereo by photometric ratios. Journal of the Optical Society of America A 1994;11(11):3069–3078.

yong Noh J,Neumann U. Expression cloning. SIGGRAPH'01; 2001. p. 277–288.

Yu T, Xu N, Ahuja N. Shape and view independent reflectance map from multiple views. International Journal of Computer Vision 2006;73(2):123–138.

Zhang L, Curless B, Seitz S. Spacetime stereo: shape recovery for dynamic scenes Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 2; 2003. p. II–367–374.

Zhang L, Snavely N, Curless B, Seitz SM. Spacetime faces: high-resolution capture for modeling and animation. ACM Annual Conference on Computer Graphics; 2004. p. 548–558.

Zhang S, Yau ST. High-speed three-dimensional shape measurement system using a modified two-plus-one phase-shifting algorithm. Optical Engineering 2007;46(11):113–603.

Zhang S, Li X, Yau ST. Multilevel quality-guided phase unwrapping algorithm for real-time threedimensional shape reconstruction. Applied Optics 2007;46(1):50–57.

Zhao X, Dellandrea E, Chen L, Kakadiaris IA. Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model. IEEE Transactions on Systems, Man, and Cybernetics, Part B 2011;41(5):1417–1428.