Reconstructing Complex Surfaces
from
Multiple Stereo Views

P. Fua
SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025, USA
(fua@ai.sri.com)

Abstract

We present a framework for 3–D surface reconstruction that can be used to model fully 3–D scenes from an arbitrary number of stereo views taken from vastly different viewpoints. This is a key step toward producing 3–D world-descriptions of complex scenes using stereo and is a very challenging problem: real-world scenes tend to contain many 3–D objects, they do not usually conform to the 2-1/2–D assumption made by traditional algorithms, and one cannot take it for granted that the computed 3–D points can easily be clustered into separate groups. Furthermore, stereo data is often incomplete and sometimes erroneous, which makes the problem even more difficult.

By combining a particle-based representation, robust fitting, and optimization of an image-based objective function, we have been able to reconstruct surfaces without any *a priori* knowledge of their topology and in spite of the noisiness of the stereo data.

Our current implementation goes through three steps—initializing a set of particles from the input 3–D data, optimizing their location, and finally grouping them into global surfaces. Using several complex scenes containing multiple objects, we demonstrate its competence and ability to merge information and thus to go beyond what can be done with conventional stereo alone.

1 Introduction

Reconstructing arbitrary 3–D surfaces from stereo image pairs—or more generally N-tuplets—remains an unsolved problem because they provide incomplete, and sometimes erroneous, information about the location of 3–D points in space. Grouping these points into meaningful surfaces involves solving a problem akin to the notoriously hard segmentation problem. As a result, a majority of recent computer vision approaches to surface reconstruction rely on much cleaner sources of 3–D data—such as laser range maps or medical volumetric data—as their input. Many of these approaches also assume that there is one, and only one, object of interest whose topology is known in advance so that a particular model—be it rigid or deformable—can be fit to the data.

Stereo, however, is a purely passive technique that only involves the use of relatively inexpensive and dependable sensors and often is the only feasible approach. We propose a novel alternative to conventional stereo reconstruction that overcomes many of the difficulties that traditional approaches encounter in the reconstruction of 3–D surfaces from stereo imagery.

More specifically, to recover surfaces from stereo, one must contend with the following problems:

- Real-world scenes often contain several objects whose topology may not be known in advance: some surfaces are best modeled as sheets while others are topological spheres or contain holes. One cannot typically assume that there is only one object and one surface of interest or expect to be able to easily cluster the 3–D points derived from stereo into semantically meaningful groups.

- The 2-1/2–D assumption that most traditional interpolation schemes make is no longer valid when reconstructing complex 3–D scenes from arbitrary numbers of images and arbitrary viewpoints. Surfaces often overlap and may be visible in one view but not another.

- The 3–D points derived from disparity maps form an irregular sampling of the measured surfaces. In addition, small errors in disparity can result in large errors in world position.

- Even the best stereo algorithms make occasional blunders that must be identified and eliminated. Furthermore, the corresponding erroneous 3–D points are often correlated with one another so that they cannot be eliminated by robust estimation alone.

To address these problems, our approach

- Relies on an object-centered representation that can handle surfaces of arbitrary complexity, specifically a set of connected surface patches or "oriented particles" as defined by Szeliski and Tonnesen [1992].

- Refines the surface's description by minimizing an objective function that combines terms measuring both surface smoothness and gray scale correlation in the input images of the surface points'projections.

We typically start with a set of stereo pairs or triplets. We compute disparity maps for each of them, fit local quadric surface patches to the corresponding 3–D points, and use these patches to instantiate a set of particles. We then refine their positions by minimizing the objective function, thereby returning to the original image data. Finally, we impose a metric upon the set of particles that allows us to cluster them into meaningful global surfaces.

Our technique allows the modeling of complex 3–D scenes whose topology is unknown *a priori* from stereo data without the need to rely on other sources of range data. This ability is valuable for applications, such as robotics and high-resolution cartography, where precise scene models are not always available. For example a mobile robot must be able to distinguish the ground from objects lying on it to model obstacles and avoid them. Similarly, to grasp and manipulate objects, a robot must be able to distinguish and model them. In the case of low-resolution cartography, the earth's surface can indeed be relatively well modeled as a single surface. At high-resolution, however, this stops being true: objects such as trees or buildings must be modeled as separate 3–D surfaces, and the usual 2-1/2–D assumptions break down.

In the following section, we describe related work and our contributions in this area. We then present our framework in detail, discuss the procedure's behavior on synthetic data, and show results on real images of complex scenes.

2 Related Work and Contributions

Many recent publications describe the reconstruction of a surface using 3–D object-centered representations, such as 2-1/2–D grids [Robert *et al.*, 1992], 3–D surface meshes [Cohen *et al.*, 1991, Delingette *et al.*, 1991, Terzopoulos and Vasilescu, 1991, Vemuri and Malladi, 1991, McInerney and Terzopoulos, 1993, Koh *et al.*, 1994, Chen and Medioni, 1994], parameterized surfaces [Stokely and Wu, 1992, Lowe, 1991], local surfaces [Sander and Zucker, 1990, Ferrie *et al.*, 1992, Stewart, 1994], particle systems [Szeliski and Tonnesen, 1992], spanning trees [Hoppe *et al.*, 1992], and volumetric models [Pentland, 1990, Terzopoulos and Metaxas, 1991, Pentland and Sclaroff, 1991, Park *et al.*, 1994]. Most of these rely on previously computed 3–D data, such as the coordinates of points derived from laser range finders or correlation-based stereo algorithms, and reconstruct the surface by fitting it to these data in a least-squares sense. In other words, the derivation of the 3–D data is completely divorced from the reconstruction of the surface. Errors and imprecisions in the original 3–D input data can jeopardize irretrievably the quality of the reconstruction.

In previous work, we began addressing this issue by developing an approach to 3–D surface reconstruction that uses image cues while recovering the surface's shape [Fua and Leclerc, 1994a, Fua and Leclerc, 1994b]. It uses an object-centered representation—specifically a 3–D triangulated mesh—and can take advantage of both monocular shading cues and stereoscopic cues from any number of images to refine the model while correctly handling self-occlusions.

However, in our previous approach—as in most of those mentioned above—it is assumed that the range data used to initialize the models can easily be clustered into separate groups that define distinct objects. A separate 3–D model can then be fitted to each object. For complex scenes, this clearly is much too strong an assumption. The technique presented in this paper, was specifically designed to eliminate the need for this assumption. It replaces our triangulated meshes by a particle system that does not require any *a priori* knowledge of the surface's topology and lends itself to the definition of a metric that has allowed us to effectively cluster local surface patches into global ones.

The particles can adjust so as to minimize an objective function that is the sum of an image-based term and of a regularization term. The image-based term is a generalization of the multi-image intensity correlation term we used in our mesh work, except for the fact that we replace the triangular facets by circular surface patches attached to each particle. Optimization allows us to refine the position of legitimate particles and to eliminate spurious ones. Consequently, we do not have to depend on the input 3–D data to be error-free to achieve good results. This is in contrast to Szeliski and Tonnesen's particles [1992] that have been used to great effect to model high quality medical data but make no provisions for noisy and incorrect data points.

To instantiate our set of particles, we robustly fit local surfaces to the raw disparity data in a manner that is closely related to other local surface techniques e.g., [Sander and Zucker, 1990, Ferrie *et al.*, 1992, Hotz *et al.*, 1993, Stewart, 1994]. In our approach, however, these local surfaces are not the end result since they can be refined and either accepted as part of a global surface or rejected. In spirit, our approach is closely related to that advocated by Hoff and Ahuja [1987] because it combines the processes of feature matching and surface interpolation. Unlike this earlier work, however, our technique is not limited to pairs of images.

We view the central contribution of this paper as twofold. First, we provide a specific alternative framework that explicitly addresses many of the problems of conventional stereo and a representation in which they can be solved. Second, we demonstrate that our implementation of this framework actually solves some of the problems involved in reconstructing complex 3–D surfaces of unknown topology and provides a foundation on which to build a solution for the others.

3 From Raw Stereo Data to Global Surfaces

Our approach to recovering complex surfaces is to model them as sets of local surface elements that interact with one another. Following Szeliski and Tonnesen [1992], we refer to the surface patches as "oriented particles." The forces that bind them can be understood as "breakable springs" [Terzopoulos and Vasilescu, 1991] that tend to align the particles with each other but may break for particles that are too far out of alignment.

We use several sets of stereo pairs or triplets of a given scene as our input data. We assume that the images are monochrome and registered so that their relative camera models are known *a priori*. Since we are interested in reconstructing surfaces, we start the process by using a simple correlation-based algorithm [Fua, 1993] to compute a disparity map for each pair or triplet and by turning each valid disparity value into a 3–D point. If other sources of range data were available, they could be used in a similar fashion. These 3–D points typically form an extremely noisy and irregular sampling of the underlying global 3–D surfaces. To effectively model these, we take the following steps:

1. Robustly fit particles to the raw 3–D points.
2. Refine the position and orientation of the particles by minimizing an image-based objective function.
3. Eliminate spurious particles and cluster those that appear to belong to the same global surfaces.

In the remainder of this section we first introduce our particles and then describe our technique in more detail.

3.1 Oriented Particles

Our surface elements are disks whose geometry is defined by the positions of their centers, the orientations of their normals and their radii. In theory, these disks have six degrees of freedom. However, when modeling a global surface in terms of such disks, translations along the tangent plane of the surface can be ignored as long as the disks remain roughly equidistant from one another and the radius can be chosen so that the disks approximately cover the surface. Therefore, in practice, we deal with only three degrees of freedom.

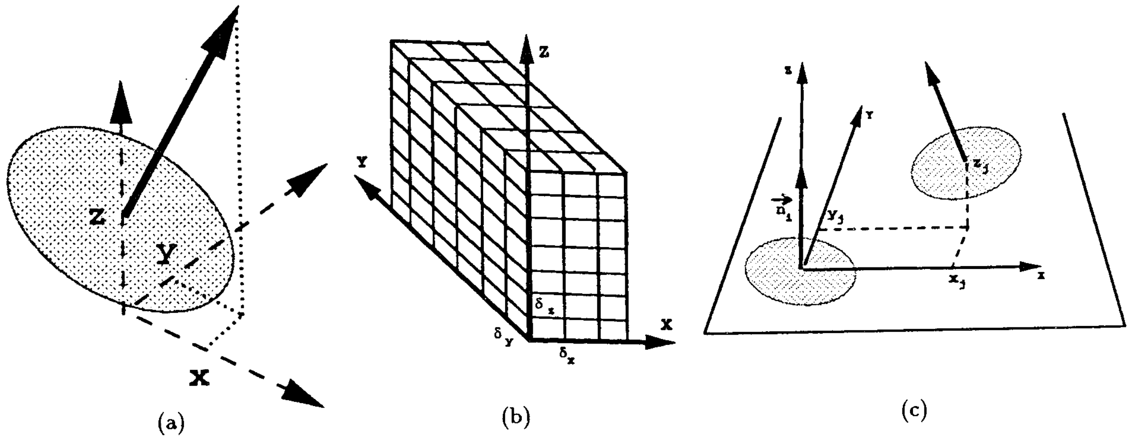

Figure 1: Data structures and metric. (a) A particle is a disk to which we associate a local referential. We allow the center of gravity to shift along the z axis and parametrize the orientation using the projections of the normal vector on the x and y axes. (b) The input 3–D points are stored in a cube-shaped set of voxels and we instantiate a particle in each voxel containing enough such points. (c) The "distance" between two particles is primarily a function of the distance of the center of gravity of one particle from the tangent plane of another.

As shown in Figure 1, to achieve an orientation-independent implementation, we assign to each particle a local referential. As discussed in Section 3.2, we instantiate this referential by robustly fitting surface patches to the 3–D points within local neighborhoods and using the surface normals to define the local z directions. We define a particle's position by the translation of the center along the local vertical and its orientation by the x and y projections of the normal on the local x and y axes. This particular parametrization is most favorable when the local vertical is relatively close to the normal of the surface under consideration for two reasons. First, the x and y projections of the particle's normal vector will then be relatively small and the interaction forces between particles almost quadratic in terms of those parameters. Second, displacements along the local z axis will be close to being normal to the underlying surface and thus precisely the ones that are most significant in terms of recovering its shape.

3.2 Initialization

To generate a set of regularly spaced particles from our noisy stereo data, we pick spatial step sizes δ_x,δ_y and δ_z along the X,Y and Z axes of an absolute referential. We use them to define a cube-shaped set of 3–D buckets, such as the one of Figure 1(b). We then store the 3–D points computed from our initial correlation data into the appropriate buckets. By fitting a local surface to every bucket containing enough points, we generate particles whose center is the projection of the bucket's center onto the surface and whose orientation is given by the surface's normal at that point. In the presence of very noisy data, the projection may fall outside of the bucket. In this case, we reject the particle thereby, ensuring that there is only one particle per bucket and that the particles are regularly distributed.

In general, most of the 3–D buckets will be empty. Therefore we do not store the set of 3–D buckets as a cube but as a hash-table allowing for both compact storage and easy computation of neighborhood relationships.

For the initialization phase to be successful, it is important both to choose the right kind of surface model and to use a robust method to perform the fitting. We have used both planar and quadric models. The quadrics, even though they involve more computation, have proven very effective because they allow the use of larger sets of points than planes without introducing any appreciable bias. In our implementation, we take advantage of this by fitting, to each bucket containing enough points, a plane of form

$$ax + by + cz = h$$

when x,y and z are coordinates in the absolute referential. We then fit a quadric of form

$$z' = ax'x' + bx'y' + cy'y' + dx' + ez' + f$$

where x',y' and z' are coordinates defined by the plane. We use not only the points in the bucket under consideration but also in the buckets that are its immediate neighbors. This method allows us to fit local surfaces of arbitrary orientation using a relatively large set of 3–D points, and tends to enforce consistency of orientation among neighboring particles.

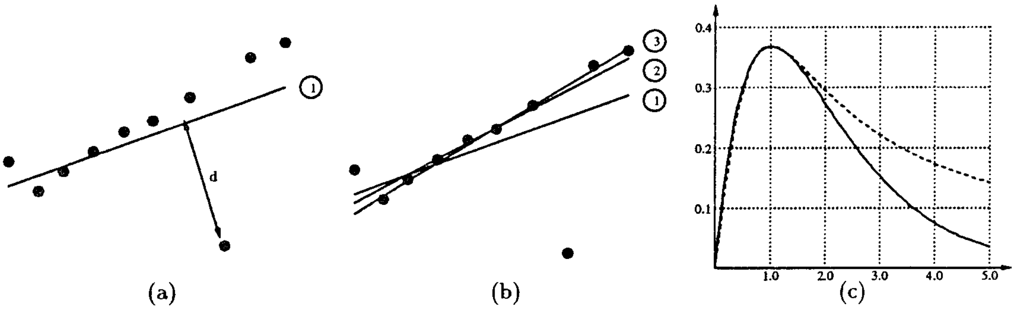

Figure 2: Iterative Reweighted Least Squares. (a) We first use standard least squares to fit an initial surface patch to the set of noisy points and compute the distance d of these data points to the fit. The points are denoted by the black circles, and the initial patch is labeled 1. (b) We weigh each data point inversely proportionally to d, fit a new surface, and iterate the procedure, thereby generating the fits labeled 2 and 3. After a few iterations, the contribution of the outliers becomes negligible. (c) The influence function of the data points is shown as a solid line. For comparison's sake, a Lorentzian influence function, scaled so as to have the same maximum, is shown as a dashed line.

Because of the noisiness of the input data, a robust surface-fitting method is essential. In this implementation, we use a variant of the Iterative Reweighted Least Squares [Beaton and Turkey, 1974] technique, as illustrated by Figure 2. To fit a surface of equation $f(x, y, z) = 0$ to a set $(x_i, y_i, z_i)_{1 \leq i \leq n}$ of n 3–D points, we minimize a criterion of the form

$$\sum_{1 \leq i \leq n} w_i f(x_i, y_i, z_i)^2 \tag{1}$$

where the w_i are weights. At the first iteration the w_i are all taken to be equal to one so that the first fit, f_0, is a regular least-squares fit. We use f_0 to compute a new set of weights

$$\begin{aligned} w_i &= exp(\frac{-|f_0(x_i, y_i, z_i)|}{\overline{d}}) \quad \text{for } 1 \leq i \leq n \\ \overline{d} &= \text{median}|f_0(x_i, y_i, z_i)|_{1 \leq i \leq n} \end{aligned} \tag{2}$$

and to fit a new surface f_1 by reminimizing the criterion of Equation 1. In effect, we use the median distance of the points to the fitted surface as an estimate of the noise variance, and we discount the influence of points that are more than a few standard deviations away. This approach can be related to the use of Lorentzian estimators [Black and Rangarajan, 1994] because the influence function of the data points, shown in Figure 2(c), has roughly the same shape. However, it requires no *a priori* estimate of the variance of the noise and involves only least-squares, and therefore fast, minimization.

By iterating this procedure a few—typically five—times, we have been consistently able to fit our quadric surfaces to noisy data, even in the presence of large numbers of outliers. Figure 3 illustrates the robustness of our algorithm.

3.3 Clustering

To cluster the isolated particles into more global entities, we define a simple "same surface" relationship $\mathcal{R}$ between particles P_i and P_j as follows:

$$P_i \mathcal{R} P_j \iff d_{\text{part}}(P_i, P_j) < \max_d \tag{3}$$

where d_{part} is a distance function and $\max_d$ a distance threshold. We could take d_{part} to be the Euclidean distance between particle centers. However, such a distance would not be discriminating enough for our purposes because it is circularly symmetric and does not take the particles'orientation into account. It has proved much more effective to define a distance function that penalizes more heavily the distance of one particle's center from the tangent plane of the other than the distance along the tangent plane. The simplest way to achieve this result is to define d_{part} as follows:

$$\begin{aligned} d_j &= kz_j^2 + (1-k)(x_j^2 + y_j^2) \\ d_i &= kz_i^2 + (1-k)(x_i^2 + y_i^2) \\ d_{\text{part}}(P_i, P_j) &= max(d_i, d_j) \end{aligned} \tag{4}$$

where x_j,y_j and z_j are the coordinates of the center of P_j in a referential whose Z direction is the normal of P_i and whose origin is the center of P_i, as shown in Figure 1(c), and k is a constant larger than 0.5. In this paper, we take $k = 0.9$; x_i,y_i and z_i are defined symmetrically.

In essence, the threshold $\max_d$ on d_{part} limits the curvature of the common underlying surface to which particles may belong. As such it is domain dependent; here we take $\max_d$ to be a multiple, typically 1.5, of the median value of d_{part} for all pairs of neighboring particles in the cube-shaped structure of Section 3.2.

The data set equipped with the relationship $\mathcal{R}$ can now be viewed as a graph whose connected components are the surfaces we are looking for. In practice, there will usually be spurious particles that are weakly linked to legitimate clusters. In such cases, we have found that removing all points that do not have a minimum number of neighbors allows us to throw away the gross errors and generate meaningful clusters such as the ones depicted in the last column of Figure 3.

3.4 Refining

Because it is extremely difficult to design a stereo algorithm that never produces correlated artifacts, we cannot expect any robust fitting technique to exclude all erroneous 3–D points. Furthermore, fitting local surfaces to the initial data amounts to smoothing and may result in spurious particles that appear to line up with legitimate ones and become very hard to eliminate.

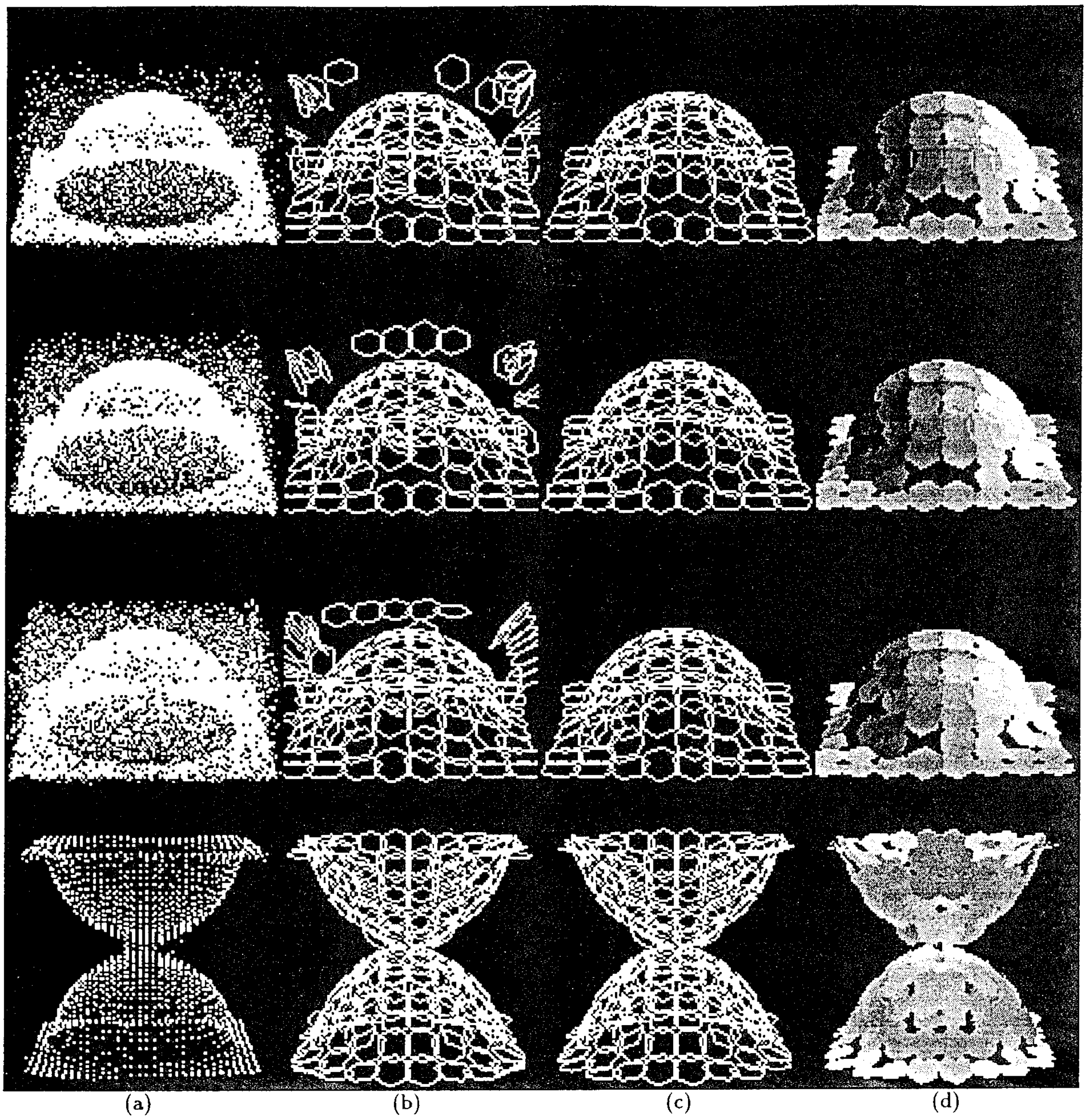

Figure 3: Initializing particle sets. (a) Four sets of noisy 3–D points. In the first three, the majority of points—90%, 80% and 60%, respectively—lie on a hemisphere while the remaining ones—10%, 20% and 40%, respectively—are uniformly randomly distributed. In the fourth set, the points lie on two hemispheres. (b) The particles instantiated by our robust fitting procedure. Note that there is not a particle in every voxel because, as explained in the text, the obviously meaningless ones are eliminated. (c) The particles that remain after clustering. All surviving spurious particles have disappeared. (d) A shaded view of the same particles. They are rendered as lambertian planar patches.

To illustrate this point and to produce a difficult example that gives rise to some of the same problems we have encountered when dealing with real imagery, we have generated the five synthetic images of a textured hemisphere occluding a plane shown in Figure 4. Our intent was to produce images ambiguous enough for a conventional correlation algorithm to compute spurious disparities and to show that our refinement procedure provides the additional power required to eliminate them.

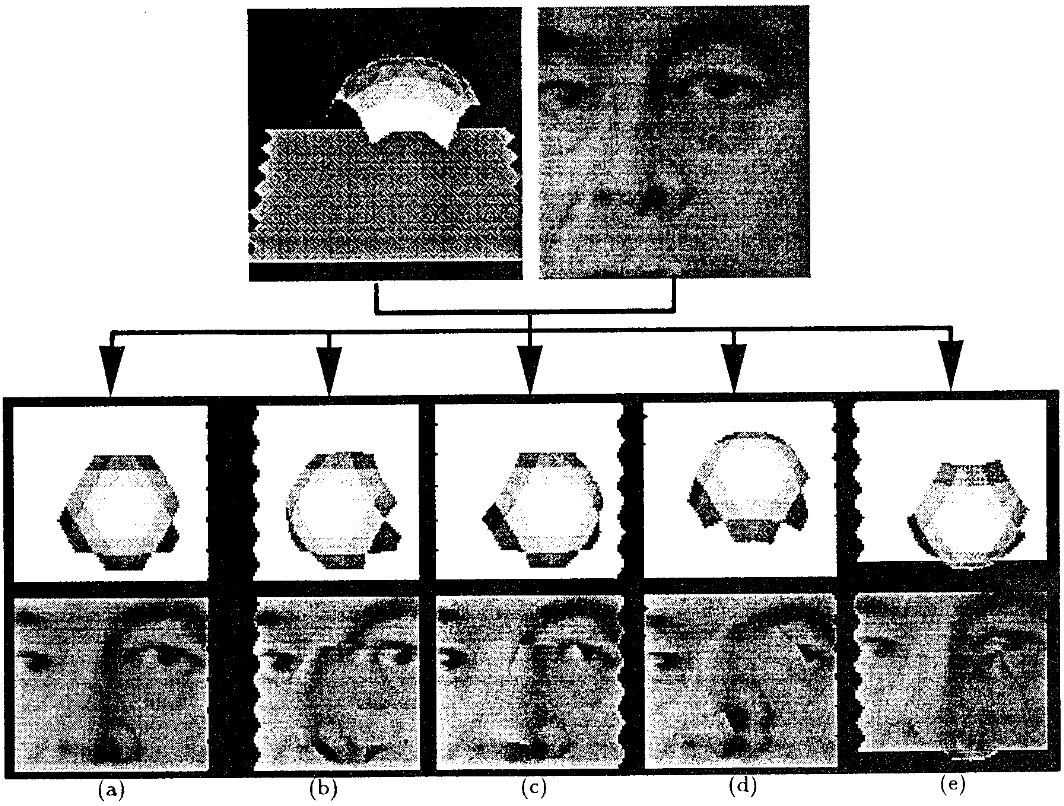

Figure 4: Constructing synthetic images of a hemisphere occluding a plane. We begin with the triangulated mesh and real face image shown in the top row. We then consider the five views depicted by the middle row and texture map the real image onto the triangulated mesh to produce the five synthetic images of the bottom row. Obviously, these five images would have been easier to parse had we mapped a different texture on the plane and the hemisphere. However, our intention in generating these images was to produce a difficult example that gives rise to some of the same problems we have encountered when dealing with real imagery.

On the left of the first row of Figure 5, we show a noise-free set of 3–D points that belong to either the plane or the hemisphere and, on the right side, the particles that are instantiated by running our initialization procedure on this set. In the leftmost column of the following rows, we show three subsets of all the 3–D points generated by correlating pairs of these images—specifically (a) and (b), (a) and (c), (a) and (d), (a) and (e)—using three different window sizes, 5x5 for the first row, 10x10 for the second, and 15x15 for the third. The second column from the left depicts the particles that are instantiated using this data.

As the window size increases, the 3–D points appear to be smoother but, in fact, fit the data less well. Of course this problem can be alleviated by using variable-shaped windows [Nishihara, 1984, Devernay and Faugeras, 1994]

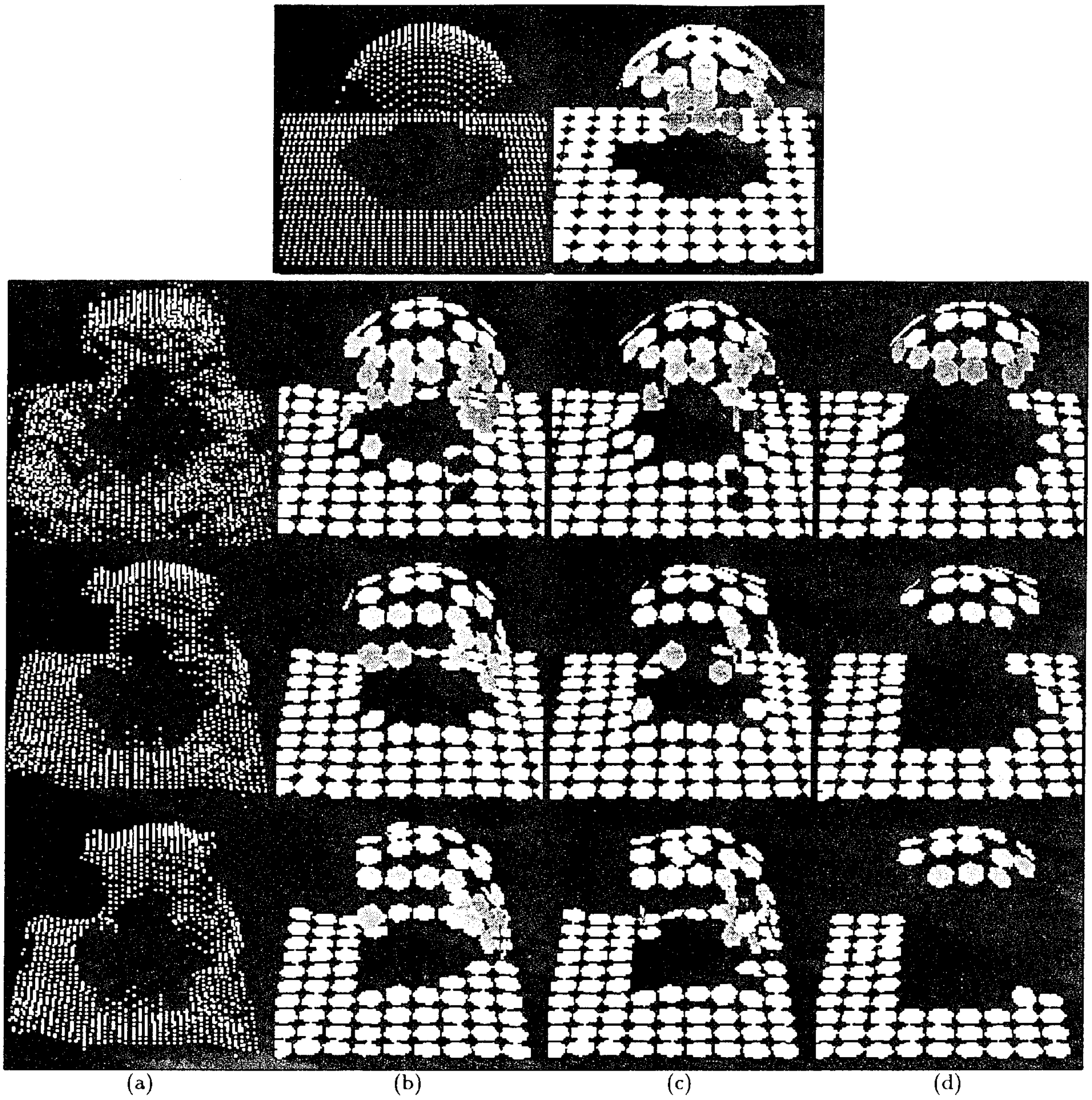

Figure 5: Running our complete procedure on the images of Figure 4. Top row: A set of 3–D points lying on the surfaces used to generate the images and a shaded view of the particles that our procedure fits to these points. Column (a): 3–D points computed by running a correlation-based algorithm on the synthetic images by using, from top to bottom, fixed windows of size 5x5, 10x10, and 15x15. Column (b): The particles instantiated from these sets of points. Column (c): The same particles after optimization. Column (d): For each set of particles, the two subsets that appear to belong to the same surface according to our metric.

and, our approach to particle refinement can be viewed as a 3–D generalization of these purely image-based techniques.

A more serious problem stems from the presence of correlated but erroneous 3–D points on the right side of the shape. These points are produced by the correlation windows that straddle the hemisphere and the plane, and line up well enough so that the fitting procedure produces a set of surface patches that appear to be valid but do not correspond to any physical surface.

To resolve such problems, it is necessary to return to the original images and assess the quality of each particle. For each disk-shaped particle, we define a "multi-image intensity correlation" term by projecting the 3–D disks into 2–D elliptical patches in each image and measuring how well these patches correlate. This term is similar to the one we defined in previous work for 3–D triangular facets [Fua and Leclerc, 1994a]. It is a function of the three degrees of freedom of each particle and can therefore be used to perform optimization.

When using noise free synthetic images such as the ones of Figure 4, we have verified experimentally that, for legitimate particles, the minimum of this image-correlation term occurs for positions and orientations that are very close to the ones computed by simply fitting the disparity maps while the position of spurious particles may be arbitrarily far from the minimum. In such an ideal case, the good particles could be separated from the bad ones on that basis alone. However, when dealing with real images that are never entirely noise free, the situation becomes more complex and the minima for the individual particles may shift substantially because of the noise. A more practical approach is then to allow the particles to interact with one another and to rearrange themselves to minimize an energy term that is the sum of the multi-image intensity correlation term discussed above and of a deformation energy term that tends to enforce consistency between neighboring particles [Szeliski and Tonnesen, 1992]. As illustrated by the two rightmost columns of Figure 5, while optimizing the energy term, the particles that actually correspond to the same underlying global surfaces will "stick together" and the ones that do not will tend to move in separate directions, stop lining up with each other and be easily eliminated by the clustering technique of Section 3.3.

Formally, we write the total energy of a set of particles $\mathcal{E}_T$ as

$$\mathcal{E}_T = \mathcal{E}_{St} + \lambda_D \mathcal{E}_D \quad , \tag{5}$$

where $\mathcal{E}_{St}$ is the multi-image intensity correlation term, $\mathcal{E}_D$ the deformation energy term, and λ_D a weighting coefficient.

We now turn to the formal definition of these energy terms.

3.4.1 Multi-Image Intensity Correlation

The basic premise of most correlation-based stereo algorithms is that the projection of the 3–D points into various images, or at least band-passed or normalized versions of these images, must have identical gray levels. To take advantage of this property in our particle-based representation, we define the stereo component of our objective function as the variance in gray-level intensity of the projections in the various images of a given sample point on a particle, summed over all sample points, and summed over all particles. This component is depicted by Figure 6(a) and is presented in stages below.

First, we define the sample points of a disk-shaped particle P, of center $\mathbf{x}_0$, radius R and normal $\vec{n}$, by noting that all points on a particle can be written as

$$\mathbf{x}_{r,\theta} = \mathbf{x}_0 + r\cos(\theta)\vec{v_x} + r\sin(\theta)\vec{v_y} \text{ for } 0 \leq r \leq R \text{ and } 0 \leq \theta \leq 2\pi \quad , \tag{6}$$

where $\vec{v_x}$ and $\vec{v_y}$ are unit vectors chosen so that $\vec{v_x}$,$\vec{v_y}$ and $\vec{n}$ form an orthonormal basis. We obtain our regularly spaced sample points $\mathbf{x}_{r_i,\theta_j}$ by taking

$$\begin{aligned} r_i &= i\frac{R}{n_r} \quad 1 \leq i \leq n_r \\ \theta_j &= j\frac{2\pi}{n_\theta(r_i)} \quad 1 \leq j \leq n_\theta(r_i) \end{aligned}$$

where $n_\theta(r_i) = 2\pi\frac{r_i}{R}n_r$ so as to ensure approximately uniform sampling of the disk.

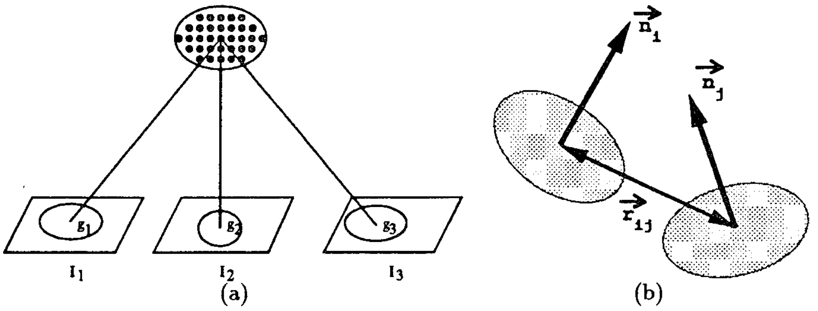

Figure 6: Computing the energy terms. (a) Particles are sampled uniformly; the circles represent the sample points. The stereo component of the objective function is computed by summing the variance of the gray level of the projections of these sample points, the g_is. (b) The forces that bind the particles are computed as functions of their normal vectors, $\vec{v_i}$ and $\vec{v_j}$, and of the vector $\vec{r_{ij}}$ whose endpoints are the centers of gravity.

Next, we develop the sum of squared differences in intensity from all images for a given point $\mathbf{x}$. A point $\mathbf{x}$ in space is projected into a point $\mathbf{u}$ in image g_i via the perspective transformation $\mathbf{u} = \mathbf{m}_i(\mathbf{x})$. Consequently, the sum of squared differences in intensity from all the images, $\sigma^2(\mathbf{x})$, is defined by

$$\begin{aligned} \mu(\mathbf{x}) &= \frac{1}{n_i}\sum_{i=1}^{n_i} g_i(\mathbf{m}_i(\mathbf{x})) \ , \\ \sigma^2(\mathbf{x}) &= \frac{1}{n_i}\sum_{i=1}^{n_i} (g_i(\mathbf{m}_i(\mathbf{x})) - \mu(\mathbf{x}))^2 \ . \end{aligned}$$

Note that $\mu(\mathbf{x})$ can be considered as the "gray level" of the sample point and by taking its median value over all sample points of the particle, we can define the median gray level of a particle that we use for clustering purposes as discussed in Section 3.5. Note also, that when using only two images, $\sigma^2(\mathbf{x})$ reduces to the square difference in gray levels. In this case, our approach is equivalent to straightforward correlation with deformable windows but has the advantage of generalizing to arbitrary numbers of images.

The above definition of $\sigma^2(\mathbf{x})$ does not take into account occlusions of the surface: not all sample points of all particles are visible in all images. In theory, this could be handled by generalizing the Z-buffering technique we used in previous work [Fua and Leclerc, 1994a]. However, in the current implementation we use a simpler heuristic. As discussed in Section 3.2, the particles are initialized by computing disparity maps using pairs or triplets of images and fitting local surfaces to the corresponding 3–D points. We record the origin of these points and associate to each particle the set of two or three images that provided the largest amount of support for the local surface. We use these images to compute $\sigma^2(\mathbf{x})$ because they are the ones in which the particle is most likely to be entirely visible.

The multi-image intensity correlation component $\mathcal{E}_{St}$ then becomes:

$$\mathcal{E}_{St} = \sum_k \mathcal{E}_{St}^k \tag{7}$$

$$\mathcal{E}_{St}^k = \frac{\sum_{r,\theta} \sigma^2(\mathbf{x}_{r,\theta})}{s_k} \ , \tag{8}$$

where s_k is the total number of samples for particle k, E_{St}^k is the value of the correlation component for a single particle and the summation over k denotes a summation over the set of all particles.

3.4.2 Particle Interactions

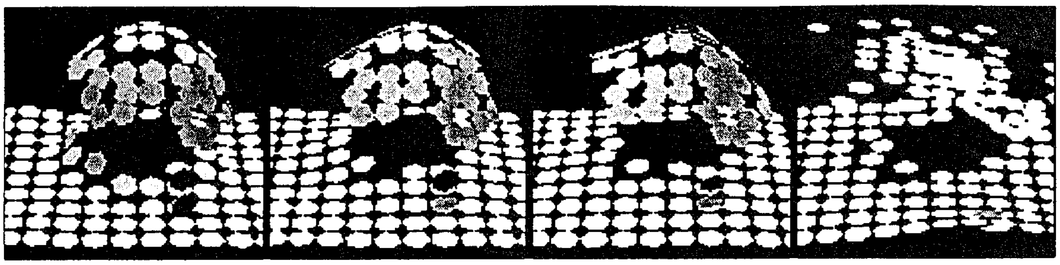

Figure 7: Rearranging the particles of the second row of Figure 5 by minimizing the deformation energy alone. The optimization progresses from left to right. All particles, but only a few outliers, end up lying on one of two separate planes.

Following Szeliski and Tonnesen [1992], we define a conormality potential $\mathcal{E}_{cn}^{ij}$ and a coplanarity potential $\mathcal{E}_{cp}^{ij}$ between particles i and j by writing

$$\begin{aligned} \mathcal{E}_{cn}^{ij} &= 1/2\|\vec{n_i} - \vec{n_j}\|^2 = 1 - \vec{n_i}\vec{n_j} \ , \\ \mathcal{E}_{cp}^{ij} &= 1/2((\vec{n_i}\vec{r_{ij}})^2 + (\vec{n_j}\vec{r_{ij}})^2) \ , \end{aligned} \tag{9}$$

where $\vec{n_i}$ and $\vec{n_j}$ are the normal vectors and $\vec{r_{ij}}$ the vector joining the centers of the two particles, as shown in Figure 6(b). These terms control the surface's resistance to bending and we take our overall regularization terms $\mathcal{E}_D$ to be

$$\mathcal{E}_D = \sum_{i,j} f(\mathcal{E}_{cn}^{ij} + \mathcal{E}_{cp}^{ij}) \tag{10}$$

where the summation over i and j denotes a summation over all pairs of particles that are neighbors in the cube-shaped structure of Section 3.1, and f is a monotonically increasing function. In practice we implement the concept of breakable springs by taking f to be

$$f(x) = log(1 + x/s)$$

with s being a fixed constant so that, as in Section 3.2, the interaction forces have a Lorentzian behavior [Black and Rangarajan, 1994]. As the particles move out of alignment, the strength of the interaction increases up to a point after which the interaction strength decreases and eventually vanishes.

In our implementation, we have not found it necessary to weight the interactions: by construction, our particles tend to be equidistant and cannot slide along the surfaces because they have only three degrees of freedom. In Figure 7, we show the result of rearranging one of the sets of particles from Figure 5 by minimizing $\mathcal{E}_D$ alone.

3.5 Global Optimization

Recall that the total energy of a set of n particles is written as

$$\mathcal{E}_T = \mathcal{E}_{St} + \lambda_D \mathcal{E}_D \ .$$

Since each particle has three degrees of freedom, $\mathcal{E}_T$ is a function of $3n$ state variables. In this work, we use conjugate gradient to minimize it and dynamically compute λ_D so that the two energy terms have comparable influences [Fua and Leclerc, 1994b].

In general, $\mathcal{E}_{St}$, the image-correlation term, is not convex and conjugate gradient may not find a global minimum. The presence of $\mathcal{E}_D$, the deformation energy, which tends to convexify the energy landscape alleviates the problem but may prove insufficient for large sets of particles. In such cases, an effective way to achieve a desirable minimum of the objective function is to cluster the particles into smaller subsets, to optimize each subset independently, and to reiterate the process until a stable solution is found. In practice, this can be achieved either by using the metric of Section 3.3 to break the set of particles into smaller subsets or by histograming the median gray levels of the particles, as defined in Section 3.4.1, and grouping those that fall into the same histogram peaks. The latter makes sense in the absence of surface markings because particles that have similar gray-levels are more likely to belong to the same underlying surfaces than particles that do not. This heuristic has been used to produce the results of Section 4 but is clearly too simple and will need refinement in future work. Note, however, that this grouping scheme is only used as a device to speed up the optimization: it need not be perfect since no final decision is based on it.

4 Results

In this section, we demonstrate the applicability of our method on a variety of imagery.

4.1 Evaluating Components of the Approach

We first use four stereo pairs of images of a head to illustrate the effectiveness of the initialization and clustering methods of Sections 3.2 and 3.3, given relatively clean stereo data. The 512x512 images, shown in Figure 8, are part of a sequence of forty that were acquired with a video camera over a period of a few seconds by turning around the subject who was trying to stand still. Camera models were later computed using standard photogrammetric techniques at the Institute for Geodesy and Photogrammetry, ETH-Zürich.

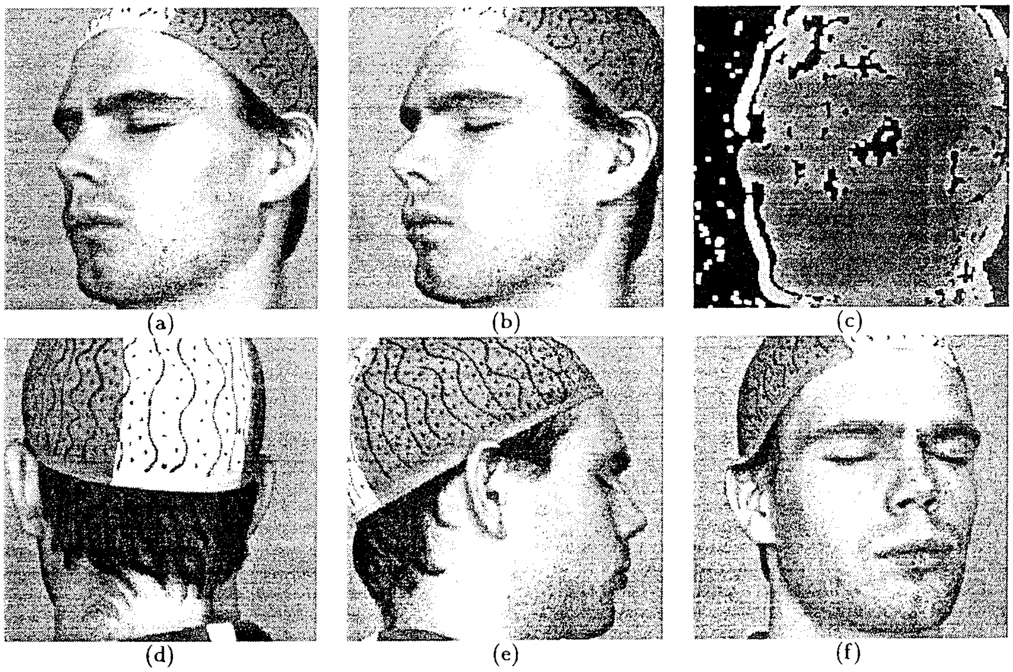

Figure 8: Head images (Courtesy of ETH-Zürich). (a,b) A stereo pair of a person's head as seen from one side. (c) The corresponding disparity map. Black indicates that no disparity value was computed, and lighter areas are further away than darker ones. Note that the disparities around the occluding contour on both the left and right sides of the head are erroneous. (d,e,f) The left images of three other stereo pairs of the same person taken from different viewpoints.

We ran our correlation-based algorithm [Fua, 1993]—once for each consecutive pair of images in the sequence—stored the resulting 3-D points in a 80x80x80 set of voxels and instantiated particles in all voxels containing at least 200 points. The results are shown in the first row of Figure 9. Because we use a large number of images, the main features of the head—including the nose, mouth, chin, ears and even the boundary of the skullcap—are

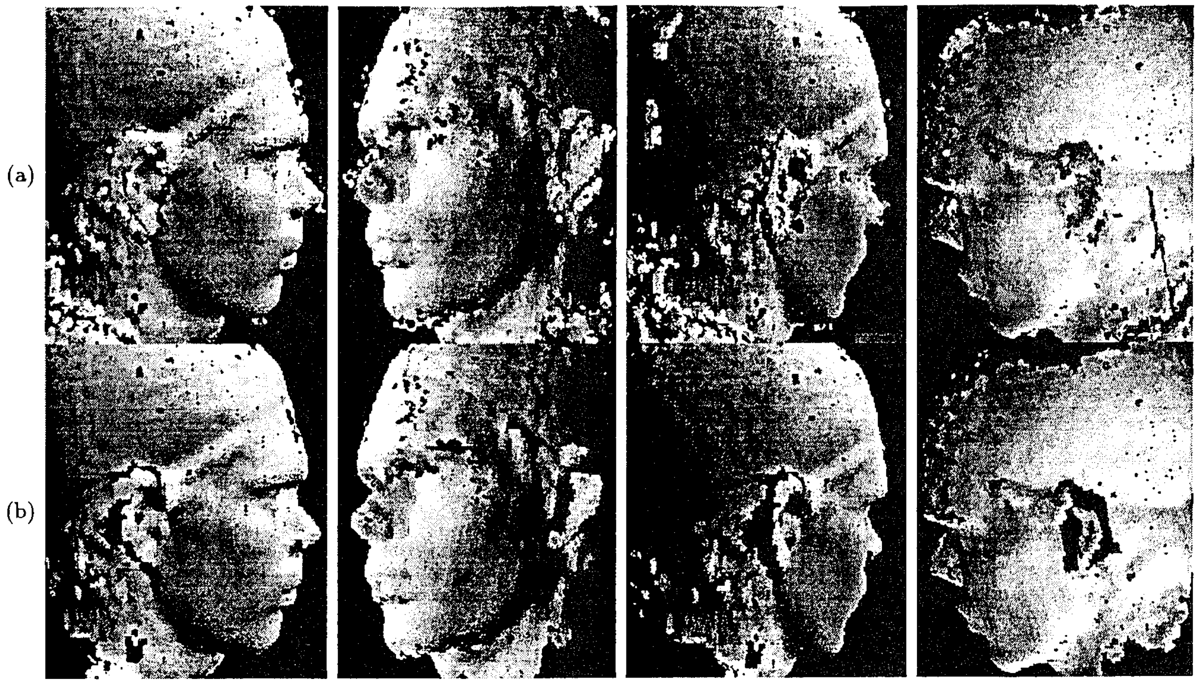

Figure 9: Modeling the head. Row (a): Four shaded views of the particles instantiated by fitting local surfaces to the 3–D points derived by correlating the images of Figure 8. Row (b): Similar views of the subset of particles that belong to the same global surface. Erroneous ones on the side of the nose and the back of the head have been eliminated.

clearly captured by our representation. However, because the correlation-based algorithm produced erroneous, but not random, disparities around occlusion boundaries, we also find a number of spurious particles around the nose, chin and back of the head. To get rid of them we computed the distance of Section 3.3 and eliminated all particles not having at least four neighbors within 1.2 times the median distance between neighbors, as shown in the second row of Figure 9. In this specific example, optimizing the positions of the particles yields a result that is virtually indistinguishable from the one presented here.

To demonstrate the accuracy that can be obtained by minimizing the total energy of Section 3.4, we use the 512x256 aerial images of Figure 10 and evaluate our results against the "ground truth" supplied to us by a photogrammetrist from Ohio State University. We ran our correlation-based algorithm, instantiated a set of particles using a 50x50x1 set of voxels, and refined their positions by minimizing the total energy of Equation 5. In Figure 10(h), we plot the vertical distance of the particles'centers from the triangulated surface defined by triangulating the control points. The Root Mean Square distance is 0.19 meter, which corresponds to an error in measured disparity that is smaller than half a pixel. Given the fact that the control points are not necessarily perfect themselves, this is the kind of performance one would expect of a precise stereo system [Güelch, 1988]. As discussed previously, in this simple case where only two images are used, our approach is equivalent to a correlation-based approach with deformable windows.

4.2 Reconstructing Complex 3–D Scenes

We now turn to the scene of Figure 11 whose modeling requires multiple viewpoints and a full 3–D representation. These images were acquired by setting a wheel and a box on a turntable and using the INRIA trinocular stereo rig to take seven triplets of images by rotating the turntable. In this case and in all views, the wheel presents surfaces

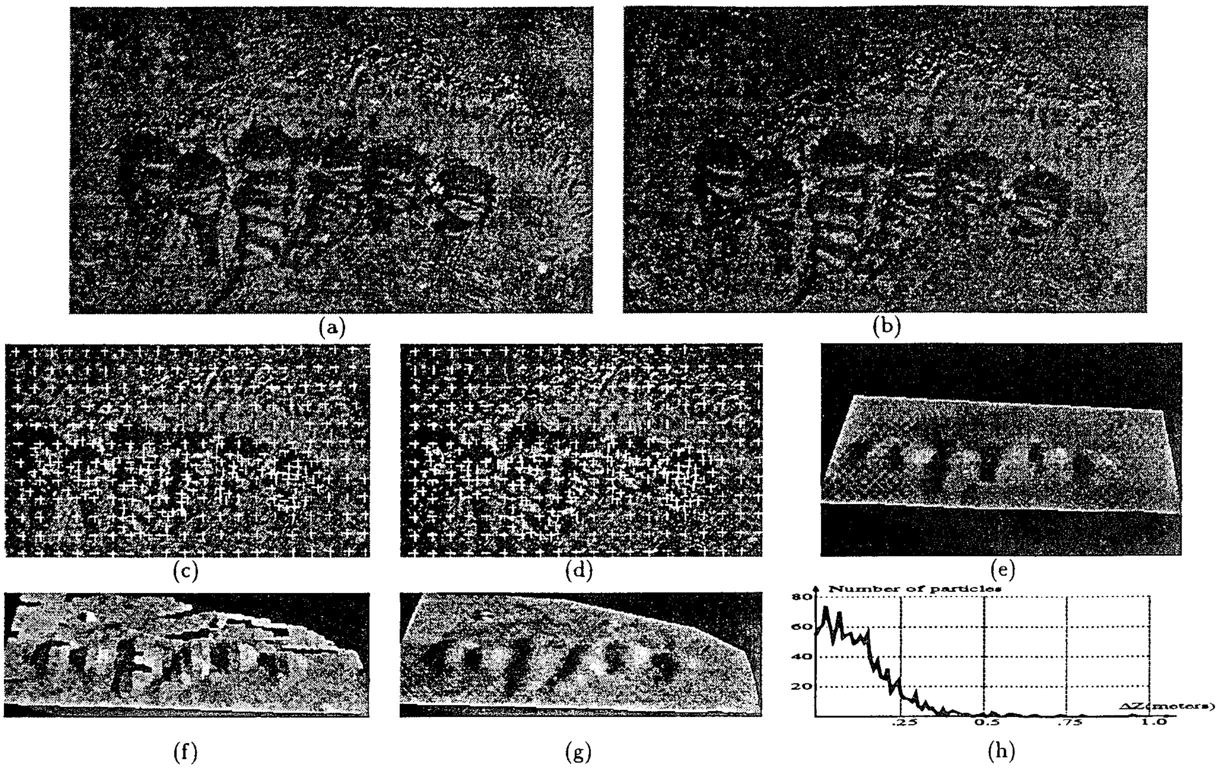

Figure 10: A stereo data set (courtesy of Ohio State University). (a,b) An aerial stereo pair. (c,d) Matched pairs of points hand-entered by a photogrammetrist. (e) Shaded view of the triangulated surface formed by the corresponding 3–D points. (f) The particles instantiated using a correlation-based disparity map after minimization of the total energy of the set. (g) Shaded view of the surface generated by triangulating the centers of the particles. (h) A plot of the vertical distance, in meters, from these centers to the "ground truth" surface of Figure 10(e). The RMS distance is equal to 0.2 meter, which corresponds to an error of approximately 0.4 pixel in disparity.

that are sharply slanted away from the cameras. The correlation data and the corresponding 3–D points, depicted by Figure 12(a), are much noisier than before. As a result, as shown in Figure 12(b), many more spurious particles are generated so that it becomes very difficult to distinguish the wheel from the base it rests on and from the erroneous fits. More specifically, there is no setting of the distance threshold that can cleanly separate the particles that sit on the wheel's surface from other ones. However, the optimization of the total energy of the set of particles produced enough of an improvement, shown in Figure 12(c), so that the clustering technique of Section 3.3 became able to effectively select the legitimate particles used to produce the shaded views of the figure's second row. This example demonstrates the ability of our object-centered representation to effectively pool the information from very different views to refine the representation beyond what could be done using only conventional stereo followed by robust surface fitting without reference to the original image data.

In Figure 13, we use the same setup and the same wheel as for the images of Figure 11, but we have added objects and use twelve triplets that span to a full 360 degrees of rotation of the turntable. The second row depicts the 3–D points computed by our simple correlation-based stereo algorithm. Note all the spurious points "floating" above the actual objects. For comparison's sake, we have verified that another correlation algorithm [Hannah, 1988], which is more sophisticated and is considered as one of the best ones currently available [Güelch, 1988], yields similar results. In fact, no setting of its parameters can effectively eliminate these points without also eliminating many of the real

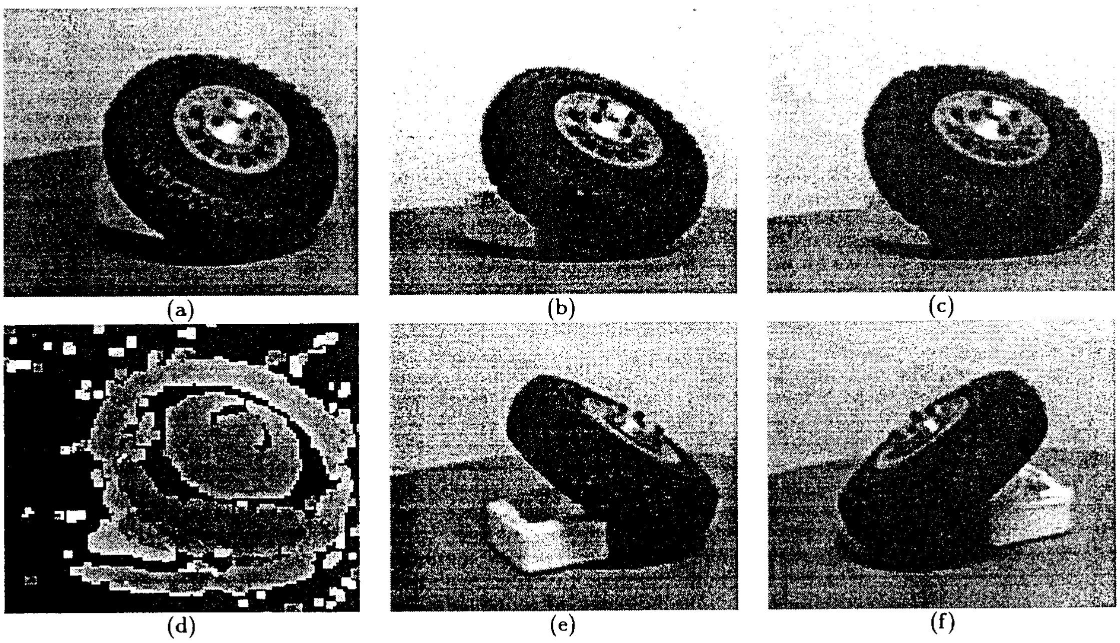

Figure 11: Images of wheel (Courtesy of INRIA). (a,b,c) A triplet of images of a wheel resting on a box sitting on a turntable. (c) Corresponding disparity map. Because the surface of the tire in the top left part of the wheel is slanted away from the camera, the disparities are of very poor quality in that part of the image. (e,f) First images of two other triplets acquired by rotating the turntable. The five other triplets used in this example were acquired for turntable positions that were intermediate between those of (e) and (f).

structures as long as one uses only pairs of images.

By running our system using the same parameters as before, except for the distance threshold, we generate the shaded representation shown in the third row of Figure 13. Note that the various objects—the wheel, the model of a brain, and the sides of the box on which they rest—appear clearly. However, our simple clustering mechanism does not pull them out as separate objects because it essentially uses the same threshold on surface curvature for the whole scene. In this case different thresholds are required for the different objects, and a more sophisticated heuristic—such as computing the distance threshold on a more local basis—would be required to achieve a complete segmentation. Finding optimal groups of particles can be recast as the problem of finding the best description of the scene in terms of this specific vocabulary given the input data. A possible approach would then be to use the Mininum Description Length [Rissanen, 1987] as was done by Leonardis *et al.* to extract surface patches from range data [1990]. Note also the hole in the wall of the wheel's tire in Figure 13(i). Careful examination of the original correlation data, reveals that 3–D data was particularly sparse there, presumably because that part of the tire is almost completely black. Such holes could potentially be filled by introducing a particle creation mechanism [Szeliski and Tonnesen, 1992].

In our final example, shown in Figure 14, we reconstruct a ground-level scene using three triplets of images acquired by the INRIA mobile robot. The ego-motion of the robot was computed by extracting 3–D segments and matching them across views [Zhang and Faugeras, 1990]. Note that the five main rocks in the scene, including one that overhangs, appear in the reconstruction.

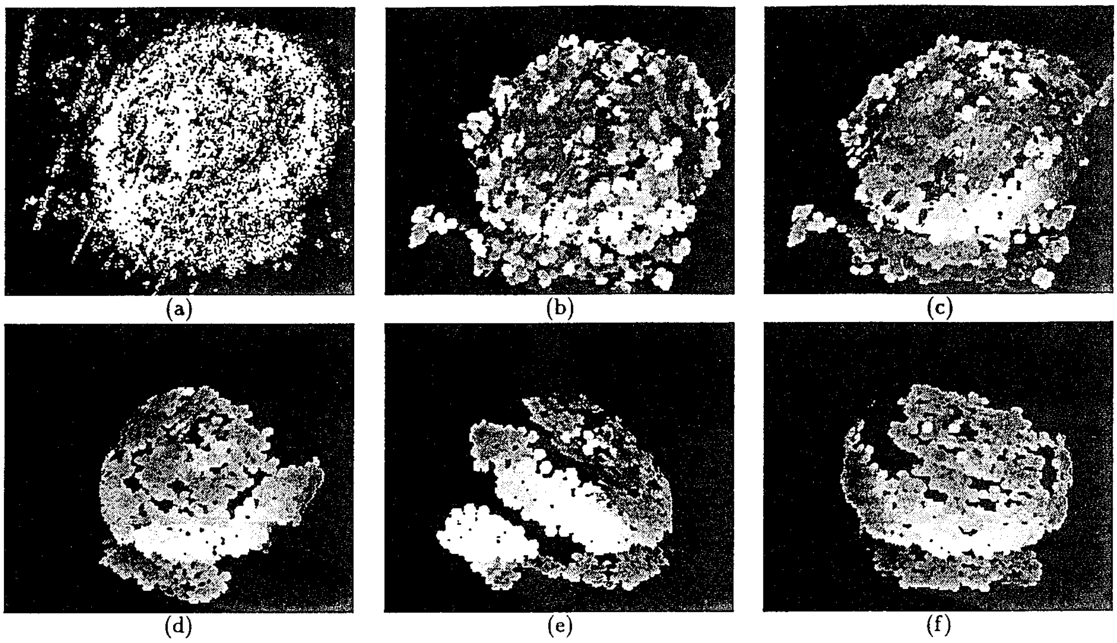

Figure 12: Modeling the wheel. (a) A subset of all the 3–D points computed by running a correlation-based algorithm on each of the image triplets. (b) The particles instantiated by fitting local surfaces to the 3–D points originating from seven noisy disparity maps. (c) The same particles after minimization of the total energy of the set. (e,f,g) Shading views of the set of legitimate particles after elimination of the spurious ones.

5 Conclusion

We have proposed a framework for 3–D surface reconstruction that can be used to model fully 3–D scenes from an arbitrary number of stereo views taken from vastly different viewpoints. By combining a particle-based representation, robust fitting, and optimization of an image-based objective function, we have been able to reconstruct surfaces without any a priori knowledge of their topology.

Our current implementation goes through three steps—initializing a set of particles from the input 3–D data, optimizing their location, and finally grouping them into global surfaces. We have demonstrated its competence and ability to merge information and thus to go beyond what can be done with conventional stereo alone.

However, as the quality of the input data decreases and the size of the problems we want to deal with increases, some of the current heuristics—specifically the ones used to cluster the particles for the purposes of both optimizing the set and rejecting outliers—may prove too simple. To push the method forward, it will be necessary to develop more sophisticated grouping techniques and to introduce new heuristics to fill in holes in the original correlation data. The basic framework, however, will remain because it provides the primitives required to implement these additional capabilities.

Acknowledgments

Support for this research was provided by various contracts from the Advanced Research Projects Agency. We wish to thank Yvan Leclerc and Richard Szeliski for their ideas—particularly the suggestion to use hash-tables to deal with neighborhood relationships—and active support in developing the approach reported here. We also wish to

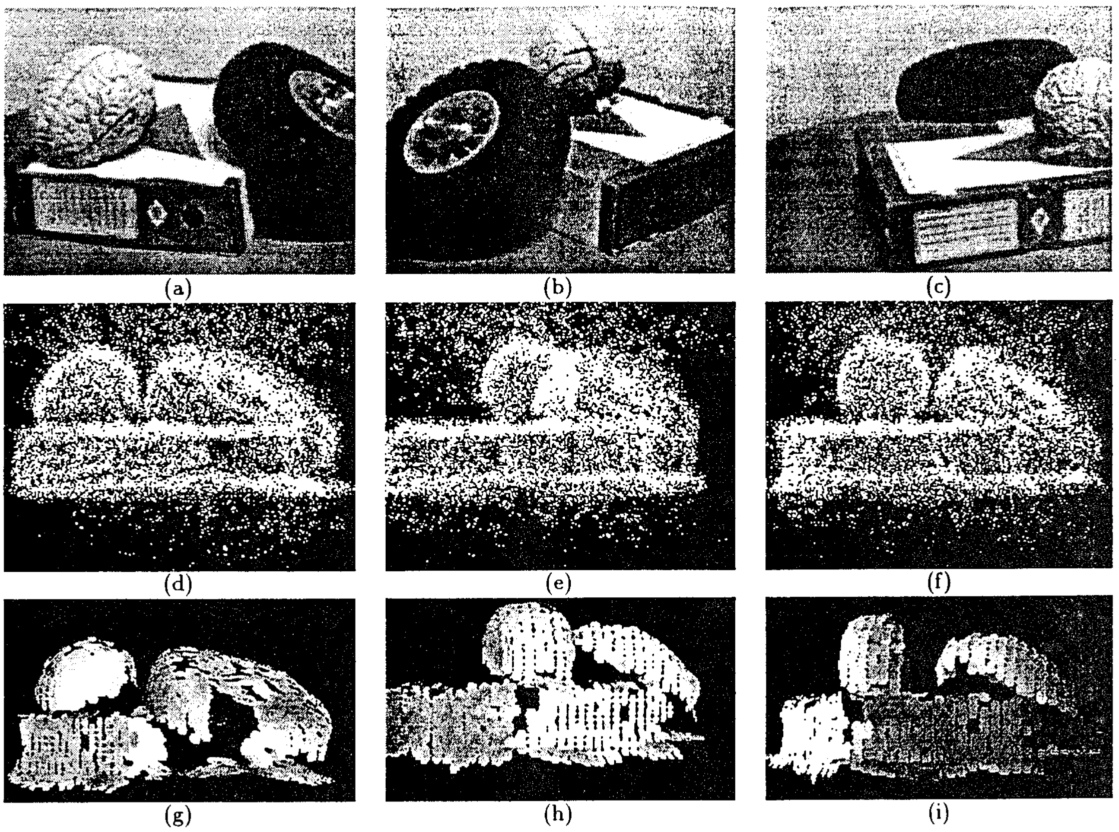

Figure 13: Images of a complex indoor scene (Courtesy of INRIA). (a,b,c) First images of three out of twelve image triplets. The three viewpoints are 120 degrees apart. (d,e,f) A subset of all the 3-D points computed by running a correlation-based algorithm on each of the image triplets. (g,h,i) Shaded views of the particle set after refinement and clustering.

thank Olivier Faugeras and the members of the INRIA Robotvis group, as well as Armin Gruën and the members of his group, who have supplied us with most of the images and calibration data that are presented here and that have proved extremely valuable to our research effort.

References

[Beaton and Turkey, 1974] A. E. Beaton and J.W. Turkey. The Fitting of Power Series, meaning Polynomials, Illustrated on Band-Spectroscopic Data. *Technometrics*, 16:147–185, 1974.

[Black and Rangarajan, 1994] M. J. Black and A. Rangarajan. The Outlier Process: Unifying Line Processes and Robust Statistics. In *Conference on Computer Vision and Pattern Recognition*, pages 121–128, Seattle, WA, June 1994.

[Chen and Medioni, 1994] Y. Chen and G. Medioni. Surface Descriptions of Complex Objects from Multiple Range Images. In *Conference on Computer Vision and Pattern Recognition*, pages 437–442, Seattle, WA, June 1994.

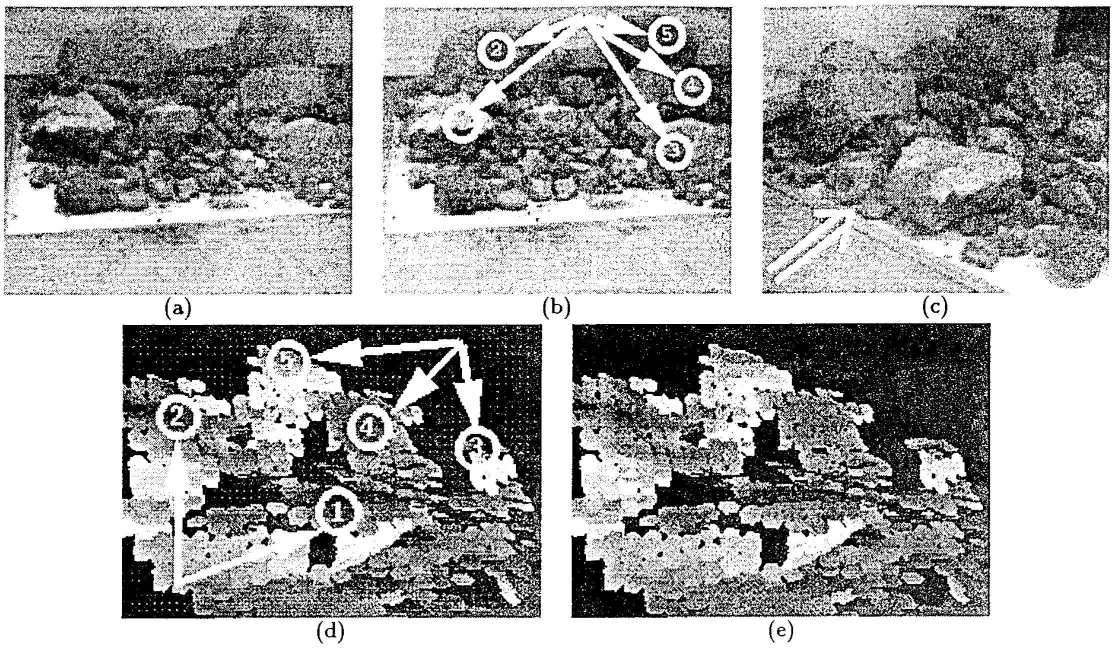

Figure 14: Modeling a pile of rocks. (a) The first image of a triplet (Courtesy of INRIA). (b) The same image with the five largest rocks labeled from 1 to 5. (c) The first image of a triplet taken from a different viewpoint. (e) Shaded view of the particle set after refinement and clustering seen from a viewpoint located on the left side of the rock pile, as indicated by the double white arrow in (c). (f) The same view with the five largest rocks labeled as in (b). Note that the overhang of rock number 1 is well recovered, a result that would be difficult to achieve using a 2-1/2-D representation.

[Cohen *et al.*, 1991] I. Cohen, L. D. Cohen, and N. Ayache. Introducing New Deformable Surfaces to Segment 3d Images. In *Conference on Computer Vision and Pattern Recognition*, pages 738–739, 1991.

[Delingette *et al.*, 1991] H. Delingette, M. Hebert, and K. Ikeuchi. Shape Representation and Image Segmentation using Deformable Surfaces. In *Conference on Computer Vision and Pattern Recognition*, pages 467–472, 1991.

[Devernay and Faugeras, 1994] F. Devernay and O. D. Faugeras. Computing Differential Properties of 3–D Shapes from Stereoscopic Images without 3–D Models. In *Conference on Computer Vision and Pattern Recognition*, pages 208–213, Seattle, WA, June 1994.

[Ferrie *et al.*, 1992] F. P. Ferrie, J. Lagarde, and P. Whaite. Recovery of Volumetric Object Descriptions from Laser Rangefinder Images. In *European Conference on Computer Vision*, Genoa, Italy, April 1992.

[Fua and Leclerc, 1994a] P. Fua and Y. G. Leclerc. Object-Centered Surface Reconstruction: Combining Multi-Image Stereo and Shading. *International Journal of Computer Vision*, 1994. In press, available as Tech Note 535, Artificial Intelligence Center, SRI International.

[Fua and Leclerc, 1994b] P. Fua and Y. G. Leclerc. Using 3–Dimensional Meshes To Combine Image-Based and Geometry-Based Constraints. In *European Conference on Computer Vision*, pages 281–291, Stockholm, Sweden, May 1994.

[Fua, 1993] P. Fua. A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features. *Machine Vision and Applications*, 6(1):35–49, Winter 1993.

[Güelch, 1988] E. Güelch. Results of Test on Image Matching of ISPRS WG III / 4. *International Archives of Photogrammetry and Remote Sensing*, 27(III):254–271, 1988.

[Hannah, 1988] M.J. Hannah. Digital Stereo Image Matching Techniques. *International Archives of Photogrammetry and Remote Sensing*, 27(III):280–293, 1988.

[Hoff and Ahuja, 1987] W. Hoff and N. Ahuja. Extracting Surfaces from Stereo Images. In *International Conference on Computer Vision*, June 1987.

[Hoppe *et al.*, 1992] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface Reconstruction from Unorganized Points. In *Computer Graphics (SIGGRAPH)*, volume 26, pages 71–78, July 1992.

[Hotz *et al.*, 1993] B. Hotz, Z. Zhang, and P. Fua. Incremental construction of local D.E.M for an autonomous planetary rover. In *Workshop on Computer Vision for Space Applications*, Antibes, France, September 1993.

[Koh *et al.*, 1994] E. Koh, D. Metaxas, and N. Badler. Hierarchical Shape Representation Using Locally Adaptative Finite Elements. In *European Conference on Computer Vision*, Stockholm, Sweden, May 1994.

[Leonardis *et al.*, 1990] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation as the Search for the Best Description of the Image in Terms of Primitives. In *International Conference on Computer Vision*, pages 121–125, Osaka, Japan, December 1990.

[Lowe, 1991] D. G. Lowe. Fitting Parameterized Three-Dimensional Models to Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(441-450), 1991.

[McInerney and Terzopoulos, 1993] T. McInerney and D. Terzopoulos. A Finite Element Model for 3D Shape Reconstruction and Nonrigid Motion Tracking. In *International Conference on Computer Vision*, pages 518–523, Berlin, Germany, 1993.

[Nishihara, 1984] H.K. Nishihara. Practical Real-Time Imaging Stereo Matcher. *Optical Engineering*, 23(5), 1984.

[Park *et al.*, 1994] J. Park, D. Metaxas, and A. Young. Deformable Models with Parameter Functions: Application to Heart-Wall Modeling. In *Conference on Computer Vision and Pattern Recognition*, pages 437--442, Seattle, WA, June 1994.

[Pentland and Sclaroff, 1991] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:715–729, 1991.

[Pentland, 1990] A. Pentland. Automatic Extraction of Deformable Part Models. *International Journal of Computer Vision*, 4(2):107–126, March 1990.

[Rissanen, 1987] J. Rissanen. *Encyclopedia of Statistical Sciences*, volume 5, chapter Minimum-Description-Length Principle, pages 523–527. John Wiley and Sons, New York, New York, 1987.

[Robert *et al.*, 1992] L. Robert, R. Deriche, and O.D. Faugeras. Dense Depth Recovery From Stereo Images. In *European Conference on Artificial Intelligence*, pages 821–823, Vienna, Austria, August 1992.

[Sander and Zucker, 1990] Peter T. Sander and Steven W. Zucker. Inferring Surface Trace and Differential Structure from 3-D Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(9):833–854, September 1990.

[Stewart, 1994] C. V. Stewart. A New Robust Operator for Computer Vision: Application to Range Data. In *Conference on Computer Vision and Pattern Recognition*, pages 167–173, Seattle, WA, June 1994.

[Stokely and Wu, 1992] E. M. Stokely and S. Y. Wu. Surface parameterization and curvature measurement of arbitrary 3-D objects: five practical methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):833–839, August 1992.

[Szeliski and Tonnesen, 1992] R. Szeliski and D. Tonnesen. Surface Modeling with Oriented Particle Systems. In *Computer Graphics (SIGGRAPH)*, volume 26, pages 185–194, July 1992.

[Terzopoulos and Metaxas, 1991] D. Terzopoulos and D. Metaxas. Dynamic 3D models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(703-714), 1991.

[Terzopoulos and Vasilescu, 1991] D. Terzopoulos and M. Vasilescu. Sampling and reconstruction with adaptive meshes. In *Conference on Computer Vision and Pattern Recognition*, pages 70–75, 1991.

[Vemuri and Malladi, 1991] B. C. Vemuri and R. Malladi. Deformable models: Canonical parameters for surface representation and multiple view integration. In *Conference on Computer Vision and Pattern Recognition*, pages 724-725, 1991.

[Zhang and Faugeras, 1990] Z. Zhang and O.D. Faugeras. Tracking and Motion Estimation in a Sequence of Stereo Frames. In L.C. Aiello, editor, *European Conference on Artificial Intelligence*, pages 747–752, Stockholm, Sweden, August 1990.