SRI International

Registration without Correspondences

Technical Note No. 537 (Revised)

August 20, 1994

By: Pascal V. Fua, Computer Scientist
Yvan G. Leclerc, Computer Scientist
Artificial Intelligence Center
Computing and Engineering Sciences Division

Approved for Public Release; Distribution Unlimited.

This work was supported in part by the Advanced Research Projects Agency under Contract No. DACA76-92-C-0008 and DACA76-92-C-0034.

The views, opinions and/or conclusions contained in this note are those of the author and should not be interpreted as representative of the official positions, decisions, or policies, either expressed or implied, of the Advanced Research Projects Agency, or the United States Government.

Registration without Correspondences

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

August 20, 1994

Abstract

In this paper, we present a method for registering images of complex 3–D surfaces that does not require explicit correspondences between features across the images. Our method relies on the use of a full 3–D model of the surface to adjust the position and orientation of the camera by minimizing an objective function based on the projections of the images onto the model. This approach constrains the camera parameters strongly enough so that the models do not need, initially, to be accurate to yield good results. When registration has been achieved, the models can be refined and the fine details recovered.

We use the 3–D surface models to adjust not only the surface's shape but also the position and orientation of the cameras by minimizing an objective function based on the projections of the model into the images. The method presented here complements our approach, described in previous publications, to the recovery of 3–D surface models from multiple images whose camera parameters are known.

Our method is applicable to the calibration of stereo imagery, the precise registration of new images of a scene and the tracking of deformable objects. It can therefore lead to important applications in fields such as augmented reality in a medical context or data compression for transmission purposes. We demonstrate its applicability by using both synthetic images and real images of faces and of terrain.

Keywords : Registration, Calibration, Surface reconstruction, Stereo,Deformable surfaces.

1 Introduction

Most of the work in recovering camera position and orientation from a set of images relies on extracting point-like features from these images. Many calibration methods rely on imaging an object whose geometry is known with great precision and exhibits features that are easy to detect; these features and their known 3–D positions can then be used to compute both the external and internal camera parameters. The works described in [Faugeras and Toscani, 1986, Tsai, 1989, Baltsavias, 1991] are examples of this approach. In the area of cartography, the calibration object is replaced by landmarks. Since landmarks or calibration grids are not always available, a large number of methods have been developed to recover relative external camera parameters, [Genery, 1979, Longuet-Higgins, 1981, Genery, 1979, Weng *et al.*, 1989, Zhang, 1993] for example, and even internal ones [Luong and Faugeras, 1993] without them. They typically use point correspondences between images and are very sensitive to errors in these correspondences, even though outlier elimination can mitigate the problem [Fischler and Bolles, 1981]. By contrast, the method proposed by Abbot and Ahuja [1990] estimates camera parameters by maximizing the agreement between surface estimates derived from several sources such as focus, stereo and vergence without the need for identifiable anchor points in the images. This method, however, assumes that a 2 1/2D representation of the scene suffices.

In this paper, we present a method for registering images of complex 3–D objects without extracting features or generating explicit correspondences and dealing with phenomena such as self-occlusions. We define registration here as the estimation of the external camera parameters for two or more images given an estimate of the object's shape and assuming that the internal parameters are known. We show that our method is capable of recovering the camera parameters even when the shape of the object is known only approximately.

We use a full three dimensional (3–D) surface models to adjust not only the surface's shape but also the position and orientation of the cameras by minimizing an objective function based on the projections of the model into the images. The projections are computed using the current estimate of the external camera parameters and the fixed internal ones.

Each point on the 3–D surface potentially projects into more than one image. In the simplest case, the objective function is the sum, over each point on the surface of the 3–D surface, of the variance of the gray levels of the point's projections in the images in which it is visible; a hidden-surface algorithm is used to account for self-occlusions. Standard optimization techniques, such as conjugate gradient descent and the simplex algorithm, are used to minimize the objective function.

The above method complements our previously described approach to the recovery of 3–D surface models from multiple images whose camera parameters are known [Fua and Leclerc, 1994b, Fua and Leclerc, 1994a]. Instead of adjusting only the surface shape to minimize the objective function as in these previous papers, here we adjust the camera parameters (and in some cases, iteratively adjust the camera parameters and surface shape). Other components of the objective function, as described in these papers, could be used in the optimization procedure. In this way we could incorporate shape-from-shading and geometric constraints. We could also include depth-from-focus information as in [Abbot and Ahuja, 1990]. However, for the sake of clarity, we describe here only the "stereo" component and refer the interested reader to the above publications.

Our method has several potential applications. In the simplest case, one can assume that the 3–D model is in fact an accurate representation of the object in the images. For example, recent work in medical imaging attempts to combine images of 3–D models derived from computer tomography (CT) and/or magnetic resonance imaging (MRI) scans with live images of the patient's shaved anatomy to assist in surgery [Lorensen *et al.*, 1993]. In this case, the 3–D model can be assumed to be accurate, and the problem is to register the 3–D model with the incoming imagery. Our approach is ideal for this situation because it does not depend on having previously defined features manually aligned with the 3–D model.

A more complex situation obtains when neither a precise 3–D model nor accurate camera models are available. For example, in the area of cartography, this may happen when a new image of a site is acquired for which only a rough elevation model (DEM) is available. In this case, we combine our surface recovery and camera parameter recovery methods into an iterative procedure: first estimate the 3–D model, then refine the camera parameters, and repeat until a stable solution is reached. Our approach offers a fully automatic procedure for block adjustment without the use of ground control points or manually designated pass points.

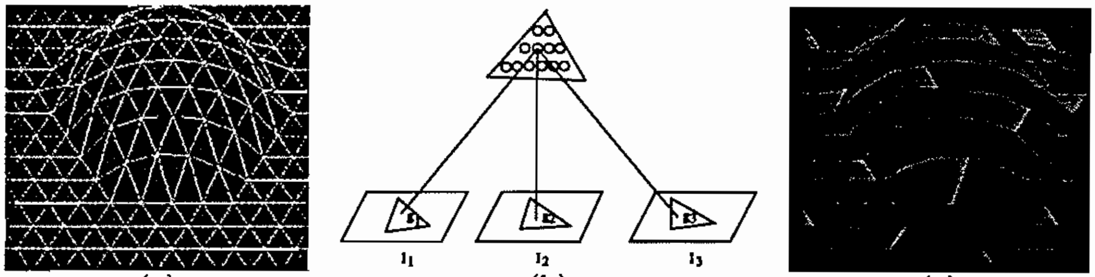
Our method's ability to recover external camera parameters as well as shape using approximate surface models is also applicable to the tracking of deformable objects. For example, a model can be acquired from an initial set of images, and can then be used to track the object's motion in subsequent frames while its shape is being iteratively adjusted.

Our iterative approach is described in Section 2 and its stability and convergence properties described in Section 3. Our experimental results, on both synthetic and real data, show that convergence can be achieved for errors in camera parameters resulting in an average absolute deviation of up to 10 pixels in the projection of the surface points into the images relative to the correct answer. Although a larger range of errors in the initial estimate of camera parameters could be accomodated if a coarse-to-fine strategy were to be used, the purpose of the experiments described here was to determine the range of errors in camera parameters that could be accomodated at any given level using a coarse-to-fine strategy. Hence the use of a dimensionless quantity, pixels, as the measure of error in the camera parameters. In Section 3, we also show that the registrations we derive are good enough for precise surface recovery. Finally, we show an application of our method to the tracking of a person whose facial expression is changing and demonstrate that the shape estimate acquired in one position is good enough to compute both the overall head motion and the deformation in the shape of the face.

2 Meshes

Our approach to recovering surface shape and camera parameters is to deform a 3–D representation of the surface so as to minimize an objective function. The free variables of this objective function are the coordinates of the vertices of the mesh representing the surface and six external parameters for each camera. The process is started with an initial estimate of the surface and camera parameters.

We represent a surface S by a hexagonally connected set of vertices called a *mesh*. The position of each vertex is specified by its x,y, and z Cartesian coordinates, and each vertex in the interior of the surface has exactly six neighbors. Neighboring vertices are further organized into triangular planar surface elements called *facets*. In Figure 1(a), we show a wireframe representation of such a mesh.



(a) (b) (c)

Figure 1: Projection of a mesh and its Facet-ID image used to accomodate occlusions during surface reconstruction: (a) A wireframe representation of the mesh. (b) Facets are sampled at regular intervals. 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. (c) The Facet-ID image, wherein the color at a pixel, shown here as a gray-level, is chosen to uniquely identify the visible facet at that point. It is used for visibility computations.

Assuming that the internal camera parameters are known, we specify the camera positions as follows. We take the position of the first one as our reference; we then specify the deviations from the initial estimate of the external parameters of the other cameras by defining a vector $\mathcal{C}$ with three rotation angles and three translations per camera.

In its most general form [Fua and Leclerc, 1994b, Fua and Leclerc, 1994a], the objective function $\mathcal{E}(\mathcal{S},\mathcal{C})$ that we use to recover the surface and camera parameters is a sum of terms that take into account the image-based constraints—stereo and shape from shading—and the externally supplied geometric constraints—features and silhouettes—that can be brought to bear on the surface. In our registration work, we have so far used only the stereo term, although we plan to use the others in the future. In the remainder of this section we first present our optimization procedure. We then describe in detail the implementation of the stereo energy term and show that it is well adapted to our registration problem.

2.1 Optimization Procedure

In theory, we could simultaneously optimize $\mathcal{S}$ and $\mathcal{C}$. However, in practice we have found it more effective to adjust them sequentially. Given an estimate of the camera parameters, we recover the surface's shape by minimizing $\mathcal{E}(\mathcal{S},\mathcal{C})$ with respect to $\mathcal{S}$. We then use the result to improve the camera parameters by minimizing $\mathcal{E}(\mathcal{S},\mathcal{C})$ with respect to $\mathcal{C}$. In Section 3, we show that this process converges in a few iterations.

We optimize the camera parameters using standard implementations of either the conjugate gradient or simplex algorithms [Press *et al.*, 1986] with very similar results.

To recover the surfaces's shape, because of the non-convexity ot the objective function, we use an optimization method that is inspired by the heuristic technique known as a *continuation method* [Terzopoulos, 1986, Leclerc, 1989] in which we add a regularization term to the objective function and progressively reduce its influence. We define the total energy of the mesh, $\mathcal{E}_T(\mathcal{S})$, as

$$\begin{aligned} \mathcal{E}_T(\mathcal{S}) &= \lambda_D \mathcal{E}_D(\mathcal{S}) + \mathcal{E}(\mathcal{S}) \ , \\ \mathcal{E}_{(}\mathcal{S}) &= \sum_i \lambda_i \mathcal{E}_i(\mathcal{S}) \ . \end{aligned}$$

The $\mathcal{E}_i(\mathcal{S})$ represent the image- and geometry-based constraints, and the λ_i their relative weights. $\mathcal{E}_D(\mathcal{S})$, the regularization term, serves a dual purpose. First, we define it as a quadratic function of the vertex coordinates, so that it "convexifies" the energy landscape when λ_D is large and improves the convergence properties of the optimization procedure. Second, in the presence of noise, some amount of smoothing is required to prevent the mesh from overfitting the data, and wrinkling the surface excessively. This is especially true when dealing with decalibrated data as shown in Figure 2.

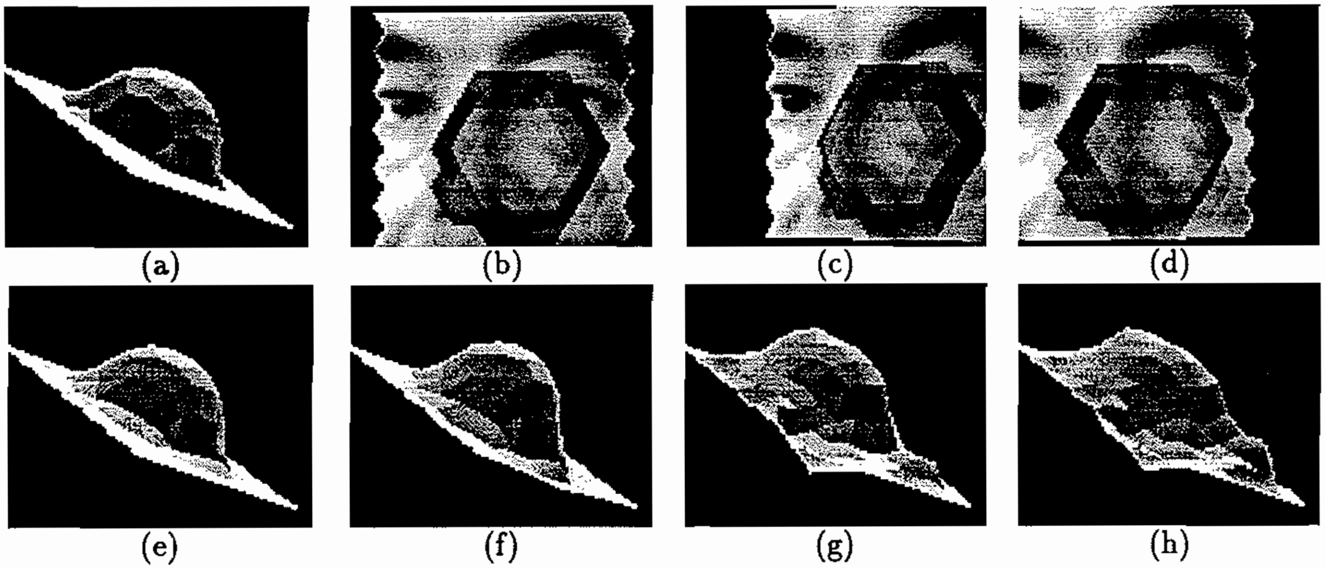


Figure 2: (a) Shaded view of a hemisphere. (b,c,d) Synthetic images generated by texture mapping the image of a face onto the hemisphere. (e,f,g,h) Recovered surface using our surface reconstruction method and progressively worse camera models. The perturbations of the camera models, as defined in Section 3.1, range from 0 to 6 pixels.

In our implementation [Fua and Leclerc, 1994b], we take $\mathcal{E}_D$ to be a quadratic term that approximates the curvature or local deviation from a plane at every vertex. It is amenable to a "snake-like" optimization technique [Kass *et al.*, 1988] in which every iteration can be reduced to solving a set of sparse linear equations. It can be shown that the dynamics of the optimization are controlled by the gradient of the objective function. As a result, we have found that an effective way to normalize the contributions of the various components of the objective function is to define a set of user-specified weights λ'_i such that

$$\sum_{1 \leq i \leq n} \lambda'_i < 1 .$$

These weights are then used to define the λs as follows:

$$\begin{aligned} \lambda_i &= \frac{\lambda'_i}{\| \vec{\nabla} \mathcal{E}_i(\mathcal{S}^0) \|}, \\ \lambda_D &= \frac{\lambda'_D}{\| \vec{\nabla} \mathcal{E}_D(\mathcal{S}^0) \|}, \end{aligned}$$

where $\lambda'_D = 1 - \sum_i \lambda'_i$ and $\mathcal{S}^0$ is the surface estimate at the start of each optimization step. Because the normalization makes the influence of the various terms comparable irrespective of actual radiometry or dimensions, we have found that the user-specified λ'_i weights are context-specific but not image-specific. In other words, we may use one set of parameters for images of faces and another when dealing with aerial images, but we do not have to change them for different faces or different landscapes.

The continuation method discussed above is implemented by taking the initial value of λ'_D to be 0.5 and then progressively decreasing it while keeping the relative values of the λ'_is constant.

In this paper, while trying to recover both camera parameters and object shape, in addition to the smoothness term, we use only the stereo term of our objective function that we denote $\mathcal{E}_{St}$ and describe below.

2.2 Stereo: 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 grey-levels. To take advantage of this property in our object-centered 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 facet, summed over all sample-points, and summed over all facets. This component is presented in stages in the remainder of this subsection.

First, we define the sample-points of a facet by noting that all points on a triangular facet are a convex combination of its vertices. Thus, we can define the sample-points $\mathbf{x}_{k,l}$ of facet f_k as

$$\mathbf{x}_{k,l} = \lambda_{l,1}\,\mathbf{x}_{k,1} + \lambda_{l,2}\,\mathbf{x}_{k,2} + \lambda_{l,3}\,\mathbf{x}_{k,3},\ l = 4, \ldots n_s,$$

where $\mathbf{x}_{k,1}$, $\mathbf{x}_{k,2}$, and $\mathbf{x}_{k,3}$ are the coordinates of the vertices of facet f_k, and $\lambda_{l,1} + \lambda_{l,2} + \lambda_{l,3} = 1$. In practice, $\lambda_{l,1}$ and $\lambda_{l,2}$ are both picked at regular intervals in $[0, 1]$. When their sum is smaller than one, $\lambda_{l,3}$ is taken to be $1 - \lambda_{l,1} - \lambda_{l,2}$. In Figure 1(b), we see an example of the sample-points of a facet.

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

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

Figure 1(b) illustrates the projection of a sample-point of a facet onto several images. The above definition of $\sigma'^2(\mathbf{x})$ does not take into account occlusions of the surface. To do so, we use a "Facet-ID" image, shown in Figure 1(c). It is generated by encoding the index i of each facet f_i as a unique color and projecting the surface into the image plane, using a standard hidden-surface algorithm.[1] Thus, when a sample-point from facet f_k is projected into an image, the index k is compared to the index stored in the Facet-ID image at that point. If they are the same, then the sample-point is visible in that image; otherwise, it is not. Let $v_i(\mathbf{x}) = 1$ when point $\mathbf{x}$ is determined to be visible in image g_i by the method above, and $v_i(\mathbf{x}) = 0$ otherwise. Then, the correct form for the sum of squared differences in intensity at a point $\mathbf{x}$ is defined by

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

When the sample-point is visible in fewer than two images (that is, when $\sum_{i=1}^{n_i} v_i(\mathbf{x}) < 2$), the above variance has no meaning and is taken to be 0. Let s_k denote the number of facet samples for facet k for which the variance is meaningful. Summing $\sigma^2(\mathbf{x})$ over all sample-points and over all facets and normalizing by the number of meaningful sample-points yields the multi-image intensity correlation component $\mathcal{E}_{St}$:

$$\mathcal{E}_{St}(\mathcal{S}) = \frac{\sum_{k=1}^{n_f} \sum_{l=4}^{n_s} \sigma^2(\mathbf{x}_{k,l})}{\sum_{k=1}^{n_f} s_k}.$$

In Figure 3 we show the result of running the stereo component of our objective function on an aerial stereo pair of a sharp ridge. We start with a coarse DEM that was provided to us by the U.S.

[1] Our algorithm is implemented on Silicon Graphics machines whose graphics hardware allows for fast computation of the Facet-ID image.

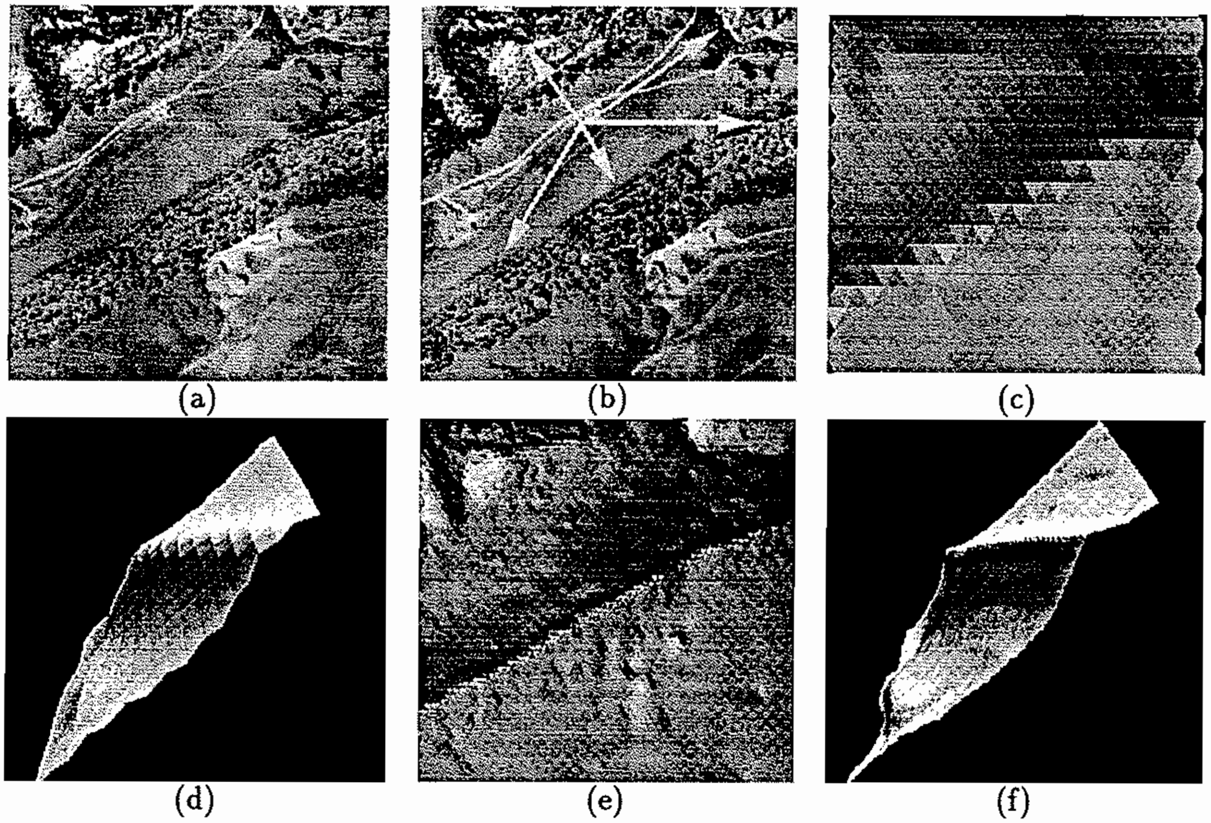


Figure 3: (a,b) A stereo pair of aerial images of a hilly site. In (b), the arrows point toward the sharp ridge in the center of the images and the shadow-casting cliffs at the top. (c) Shaded view of the triangulated DEM, seen from the viewpoint of (a). (d) Shaded view of the DEM as seen by an observer located above the upper left corner of the scene. (e,f) Shaded views of the mesh after minimization of the objective function, as seen from the viewpoints of (c) and (d). Note that the recovered ridge has become very sharp and that the shadow casting cliffs are clearly visible at the top of (e) and the bottom right corner of (f).

Geological Survey (USGS) and refine it using our method. Note that the recovered ridge, although it is seen almost edge-on, has become much sharper than in the original model result and that details in the upper part of the image are well recovered. In Figure 4, we show the reconstruction of a face using a triplet of images and adding to the stereo term the shape-from-shading term described in a previous publication [Fua and Leclerc, 1994a]. For both scenes, precise camera models have been computed using resection in one case and the INRIA calibration set-up in the other [Faugeras and Toscani, 1986]. In Section 3, we perturb these camera models to study the sensitivity of our method to errors in the camera models and its ability to remove them.

Our formulation is well adapted to tackling the registration problem because:

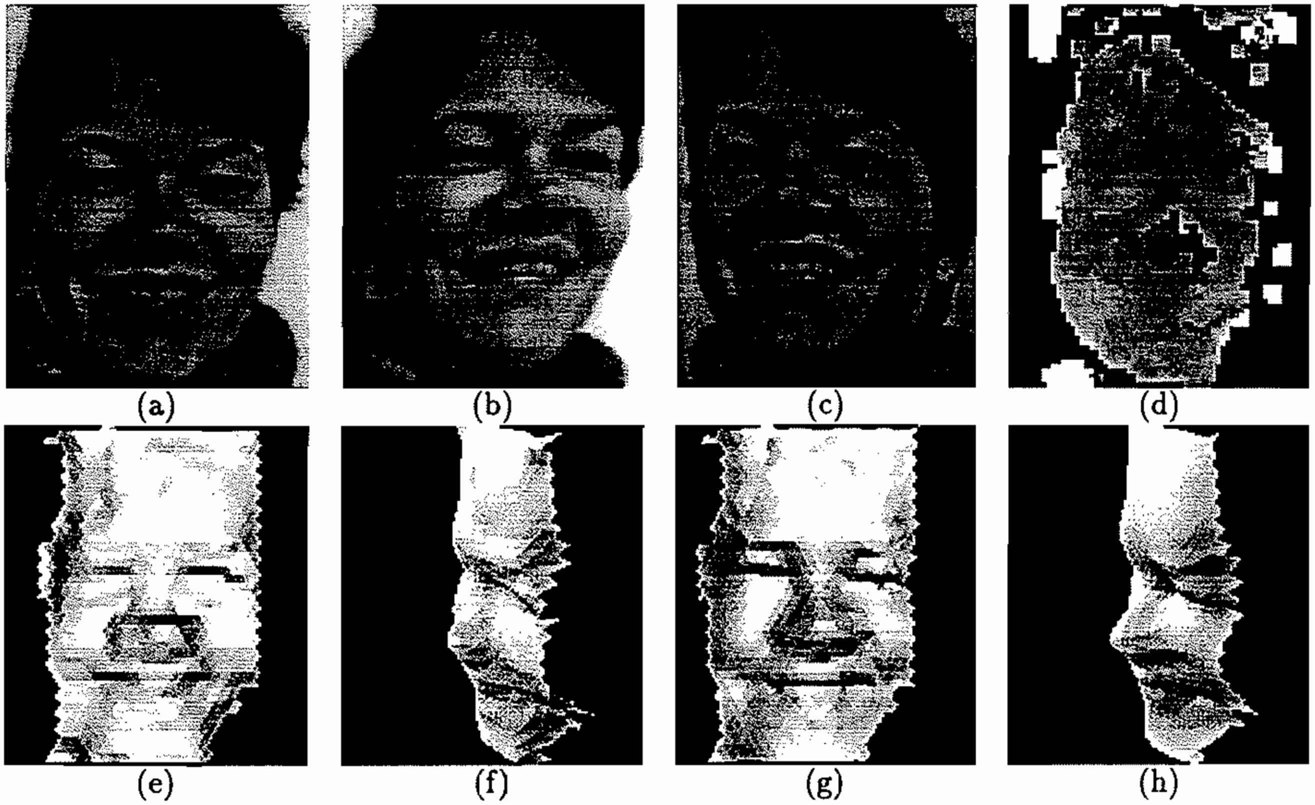


Figure 4: (a,b,c) Face images with known camera models (courtesy of INRIA) (d) Disparity map computed using a correlation-based algorithm. The black areas indicate that the stereo algorithm could not find a match. (e,f) Initial surface estimate derived by interpolating the disparity map and shown as a shaded surface. (g,h) Final surface estimate derived by combining stereo and shape-from-shading.

- It is object-centered and incorporates the use of camera models. It can therefore naturally be used to optimize both the 3–D coordinates of the surface vertices and the camera parameters.
- It can model surfaces that are slanted sharply away from the cameras. It avoids the constant depth within the window assumption of simple correlation algorithms and can therefore deal with surfaces of arbitrary orientation.
- It can deal with an arbitrary number of images, thereby strongly constraining the problem.
- It can correctly handle the self-occlusions that inevitably arise when dealing with complex surfaces.

We now turn to our experimental results.

3 Experimental Results

In this section we first quantify the ability of our method to recover external camera parameters, by perturbing the known camera models of the images of Section 2. We then show that both surface and camera parameters can be adequately recovered in the more difficult situation where the camera parameters are known only approximately and no rough estimate of the surface is initially available. Finally, we use different sets of face images to show possible applications of this technique to motion tracking and 3–D graphics.

3.1 Sensitivity Analysis

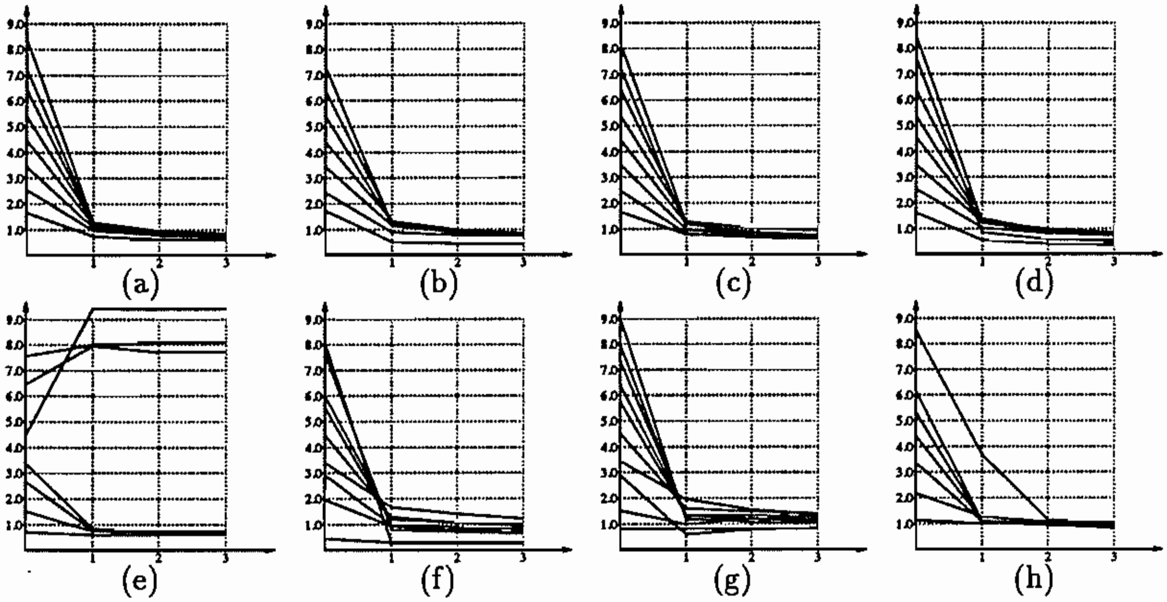


Figure 5: Perturbing and then recovering camera parameters. The meaning of the eight curves is explained in details in Section 3.1.

Here we use the synthetic images of Figure 2, the aerial images of Figure 3 and the face images of Figure 4. For each of our test scenes, we perturb the positions and orientations of the second and third cameras by random amounts. The perturbation's magnitude is defined as the average absolute deviation of the projection of the mesh vertices, measured in pixels. Each of the graphs of Figure 5 summarizes the result of one hundred trials, where a trial consists of the following steps:

- Perturb the camera models by multiplying the corresponding 3x4 projection matrix by a 4x4 translation and rotation matrix derived from six random translation and rotation values.

- Optimize the z coordinates of the mesh using the perturbed camera models and a fixed value of the regularization parameter λ'_D defined in Section 2.1.

- Optimize the camera orientation parameters using the deformed mesh.
- Reoptimize the z and external parameters twice more.

For each trial, we have recorded the initial camera perturbation and the residual perturbation after each optimization of the camera parameters. For each graph, we have grouped the trials according to the magnitude of the initial camera perturbation—between 0 and 1 pixels, 1 and 2 pixels, ...,8 and 9 pixels—and averaged the values in each group resulting in the various plots. Although a larger range of errors in the initial estimate of camera parameters could be accomodated using a coarse-to-fine strategy, the purpose of these experiments was to determine the acceptable range of errors at any given level of such a scheme. Hence the use of a dimensionless quantity, pixels, as the measure of error in the camera parameters. Ideally, all the curves should converge to zero.

The four graphs of the first row of Figure 5 were generated using the three synthetic images of Figure 2, a nosiy hemisphere as our initial surface estimate and four different values of the λ'_D regularization parameters, 0.5, 0.6, 0.7 and 0.8. Note that the curves are fairly similar in shape and converge to deviation values between 0.5 pixel and 1.0 pixel. In other words, the regularization term introduces a small bias but the method does not appear to be very sensitive to the exact amount of smoothing. We are currently investigating the use of a coarse-to-fine strategy in mesh resolution to try to improve upon this result.

The second row of Figure 5 depicts the result obtained using real images. Graph(e) corresponds to the aerial images using an approximate DEM computed by smoothing a disparity map as our initial estimate. The other three graphs were generated using the face images of Figure 4 and the rough surface model depicted in Figure 4(e). For graph(f), we used only the first two images of the triplet, for graph(g) the first and third images of the triplet, and for graph(h) all three images simultaneously. The overall behavior of the curves is the same as in the synthetic case.

In the case of the faces, the surface is complex enough to allow the recovery of camera orientation parameters, even for fairly large initial perturbations. Furthermore, when using all three images at the same time we constrain the problem more strongly and decrease the variability of the results. The curves converge towards deviation values that are around 1.0. This is due, in part, to the fact that the regularization term introduces a bias and that the camera models we use as a reference are good but not necessarily perfect; they themselves are not precise to more than 0.5 pixels. In any case, as shown in Section 3.2, the recovered registration has proved sufficient for surface reconstruction purposes.

In the case of the aerial images, however, convergence only happens for smaller values of the initial perturbation. This is not surprising in light of the results of more standard methods for computing camera parameters. It takes six matches of non-coplanar points between two images to compute the external parameters [Ballard and Brown, 1982]. Each vertex gives us a correspondence, but, in this particular case, the terrain is almost flat, except for the sharp ridge in the middle of the image. Excessive perturbation of the camera parameters is tantamount to blurring the ridge, thereby making registration based on surface point correspondences alone underconstrained.

3.2 Shape and Camera Position Recovery

Here, we demonstrate that the results shown above actually are good enough for accurate surface recovery "from scratch" when the initial camera-models are only approximately known.

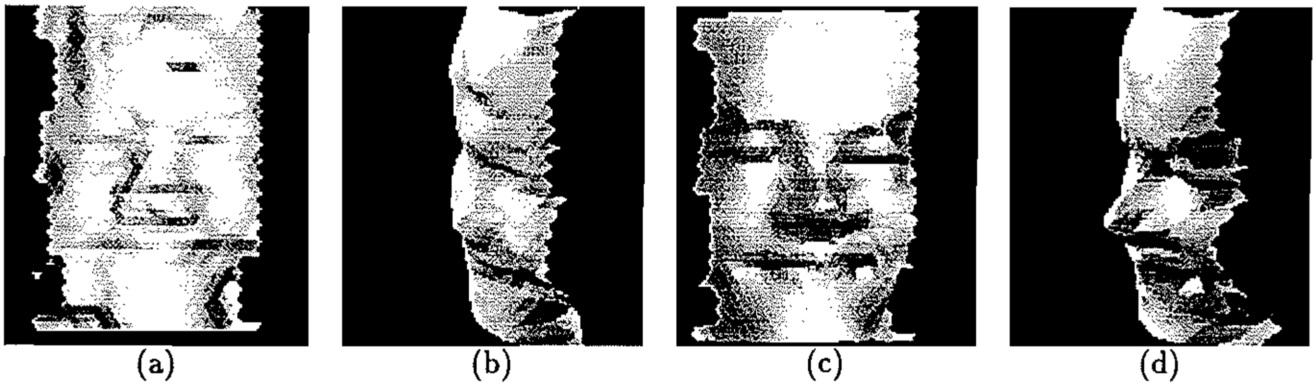


Figure 6: Recalibrating the face images of Figure 4. (a,b) Initial surface estimate derived by perturbing the camera models, recomputing a disparity map and interpolating it. Note that the mose is more flattened than previously. (c,d) Final surface estimate after registration and optimization of the initial estimates. They are almost indistinguishable from those of Figure 4.

We first use the same two scenes as above with perturbed camera models to initialize our meshes. That is, we use the epipolar lines predicted by the perturbed camera models—they are off by about four pixels—to compute the disparity maps and initial estimates shown in Figure 7(a) and Figure 6(a,b). Because of the decalibration, these initial estimates are very poor. Nevertheless, by performing the optimization of both the mesh positions and the camera parameters, we recover the surfaces shown in Figure 7(d,e) and Figure 6(c,d) that are almost indistinguishable from the ones we derived using well-calibrated data.

So far, we have shown that our method can recover camera position and orientation using approximate surface models. This capability can be used to track deformable objects, as illustrated by Figure 8, which shows face images from two triplets taken at two different times. Note that the person's facial expression has slightly changed. We use our method to detect both the global head motion between the two frames and the local deformation, as follows. We first use our standard method to recover the face in the first triplet, as shown in Figure 8(g). We then find the rotation and translation parameters that minimize our stereo objective function using the first images of both triplets and the previously derived surface estimate. Finally, we use these parameters to rotate and translate all three camera models of the second triplet and use them and the corresponding images to deform the mesh again to better conform to the new facial expression. In Figure 8(h), we show the new estimate of the surface after rotation and translation by the amount computed earlier. In Figure 8(i), we show the displacement of the vertices as a flow field from the old to the new vertex positions. Note that this flow field does not exhibit any global motion, thereby indicating that our

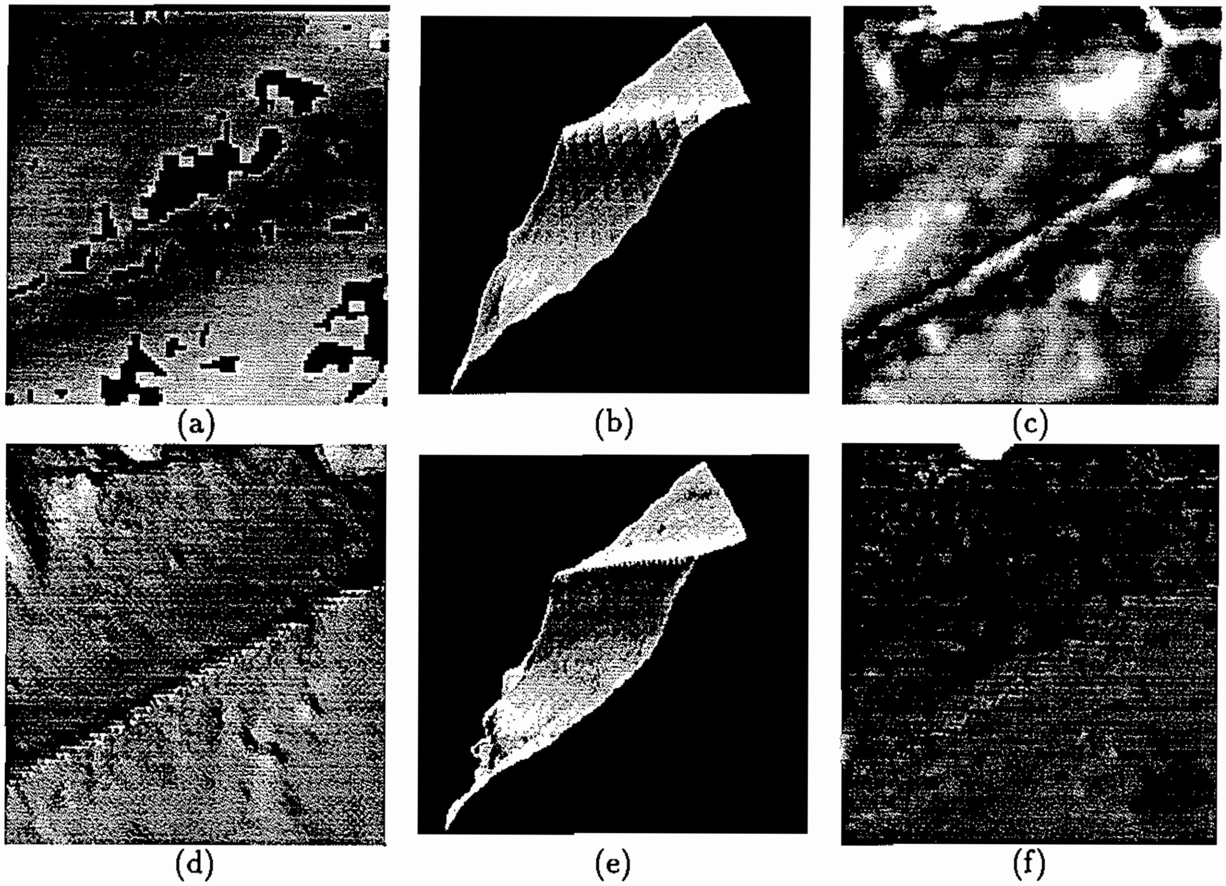


Figure 7: Recalibrating the aerial images of Figure 3. (a) Disparity map computed without camera models by assuming that the epipolar lines are horizontal. (b) Initial surface estimate derived using the disparity map and perturbed camera models. (c) Absolute differences in elevation predicted by this initial model and the optimized model depicted in Figure 3 (e,f). The image is stretched so that differences of more than 50 feet, or about two pixels in disparity, appear in white. (d,e) Shaded views of the mesh after camera parameters recovery and optimization of the mesh following the schedule used to derive the model of Figure 3. (f) Differences in elevation between the two surface estimates. Except for a small patch in the upper part of the image, they are smaller than 25 feet or one pixel in disparity.

global motion estimate is correct. This method could be generalized to the tracking of deformable objects in a video frame while updating the surface estimate. By using such an approach, the surface model can then be modified incrementally—as opposed to being recomputed at every iteration—and, since we use a physically meaningful 3-D representation, powerful methods such as Kalman

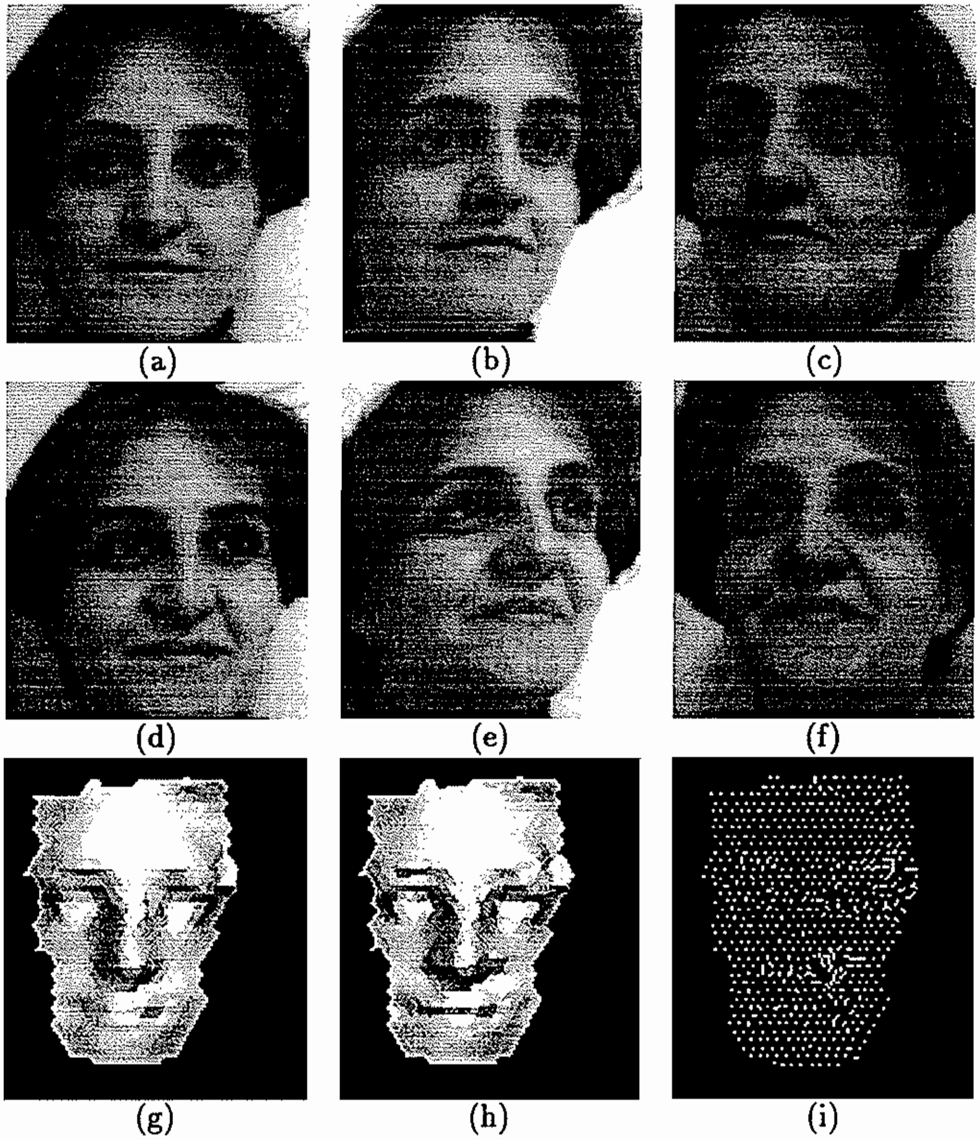


Figure 8: Application to motion tracking of deformable objects. (a,b,c) Triplets of face images taken simultaneously. (courtesy of INRIA). (d,e,f) Second triplet taken slightly later. Note that the head has moved and that the expression has changed. (g) Reconstructed surface for the first triplet. (h) Reconstructed surface for the second triplet, shown rotated and translated so that if matches the first one. (i) Flow field of the motion of the vertices of the first mesh to the rotated second one. Note that the field exhibits no overall structure and only local deformations, showing that the global motion has been correctly recovered.

filtering can be brought to bear. This technique could be applied to stabilize and transmit face images: The motion parameters would be used to perform the stabilization and one would then only transmit the parts of the face that have undergone a substantial deformation.

4 Conclusion

We have presented a method for registering images of complex 3–D surfaces that does not require explicit correspondences between point-like features across the images. Our method relies on the use of a full 3–D model of the imaged surface to recover external camera parameters. This approach constrains the camera parameters strongly enough so that the 3–D models do not need, initially, to be accurate to yield good results. Furthermore, when registration has been achieved, the models can be refined and the fine details in the surfaces of interest recovered precisely.

The method is applicable to the calibration of stereo imagery, the precise registration of new images of a scene, and the tracking of deformable objects. It can therefore lead to important applications in fields such as augmented reality in a medical context or data compression for transmission purposes.

Using static imagery, we have shown that, if the surfaces to be registered have enough relief, the method is both robust and accurate to within 1 pixel for initial errors of up to 10 pixels in camera registration. Future work will concentrate on designing a coarse-to-fine strategy to be able to handle larger errors—ideally, such as those produced by a completely decalibrated set of cameras—and on implementing a Kalman filtering style approach to the modeling of surfaces in video sequences.

Acknowledgments

Support for this work was provided by contracts from the Advanced Research Projects Agency. We wish to thank M. Fischler and O. Monga whose suggestions have greatly influenced the work described here and Hervé Matthieu for providing us with the face images and corresponding calibration data.

References

[Abbot and Ahuja, 1990] A. L. Abbot and N. Ahuja. Active surface reconstruction by integrating focus, vergence, stereo, and camera calibration. In *International Conference on Computer Vision*, pages 489–492, 1990.

[Ballard and Brown, 1982] D.H. Ballard and C.M. Brown. *Computer Vision.* Prentice-Hall, 1982.

[Baltsavias, 1991] E. P. Baltsavias. *Multiphoto Geometrically Constrained Matching.* PhD thesis, Institute for Geodesy and Photgrammetry, ETH Zurich, December 1991.

[Faugeras and Toscani, 1986] O.D. Faugeras and G. Toscani. The calibration problem for stereo. In *Conference on Computer Vision and Pattern Recognition*, pages 15–20, Miami Beach, Florida, 1986.

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

[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. Accepted for publication, 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*, Stockholm, Sweden, May 1994.

[Genery, 1979] D.B. Genery. Stereo-camera calibration. In *ARPA Image Understanding Workshop*, pages 101–107, 1979.

[Kass *et al.*, 1988] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[Leclerc, 1989] Y. G. Leclerc. Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, 3(1):73–102, 1989.

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

[Lorensen *et al.*, 1993] W. Lorensen, H. Kline, C. Nafis, R. Kikinis, D. Altobelli, and L. Gleason. Enhancing reality in the operating room. In *IEEE Visualization Conference*, pages 410–415, San Jose, California, October 1993.

[Luong and Faugeras, 1993] Q.-T. Luong and O.D. Faugeras. Self-calibration of a stereo rig from unknown camera motions and point correspondences. In *Calibration and orientation of cameras in computer vision*. Springer-Verlag, 1993.

[Press *et al.*, 1986] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes, the Art of Scientific Computing*. Cambridge U. Press, Cambridge, MA, 1986.

[Terzopoulos, 1986] D. Terzopoulos. Regularization of inverse visual problems involving discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:413–424, 1986.

[Tsai, 1989] R.Y. Tsai. Synopsis of Recent Progress on Camera Calibration for 3D Machine Vision. In Oussama Khatib, John J. Craig, and Tomás Lozano-Pérez, editors, *The Robotics Review*, pages 147–159. MIT Press, 1989.

[Weng *et al.*, 1989] J. Weng, T.S. Huang, and N. Ahuja. Motion and structure from two perspective views: algorithms, error analysis and error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(5):451–476, 1989.

[Zhang, 1993] Z. Zhang. Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. In *Conference on Computer Vision and Pattern Recognition*, pages 556–561, 1993.