Structure From Motion for Scenes Without Features[*]

Anthony J. Yezzi[‡] Stefano Soatto[§]

‡ Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta - GA 30332 ayezzi@ece.gatech.edu
§ Computer Science Department, University of California, Los Angeles - CA 90095, soatto@ucla.edu

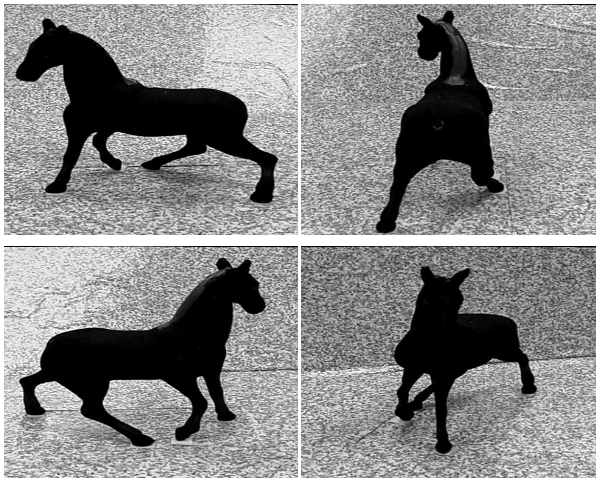

Figure 1: *Images of a scene with smooth surfaces and constant isotropic radiance. Although pictorially simple, these scenes challenge most SFM algorithms because of the lack of photometrically distinct "features".*

Abstract

We describe an algorithm for reconstructing the 3D shape of the scene and the relative pose of a number of cameras from a collection of images under the assumption that the scene does not contain photometrically distinct "features". We work under the explicit assumption that the scene is made of a number of smooth surfaces that radiate constant energy isotropically in all directions, and setup a region-based cost functional that we minimize using local gradient flow techniques.

1 Introduction

We address the problem of estimating the shape of a scene and the relative pose of a collection of cameras from a number of images. This is one of the classical problems of computer vision, known as structure from motion (SFM), and extensive literature exists, for which we refer the reader to the textbook [4]. Most of the existing work[1] in SFM concentrates on the case where the scene contains a number of photometrically distinct "features", that can be associated to geometric primitives, such as points or lines, in the scene. When this is the case, feature correspondence across images can be established in a number of ways (again, see [4] for details), at which point the problem becomes purely geometric, and the framework of epipolar geometry captures the essential relationships among corresponding points in the images and their relation to the three-dimensional structure of the scene.

In this paper, we concentrate on scenes that do not fit in this general scheme, in the sense of not having any photometrically distinct "point features", such as the scene in figure 1. We operate under the explicit assumption that the scene is composed of smooth surfaces that support a constant radiance function, which projects onto the image to yield a piecewise smooth irradiance. These scenes challenge the most common algorithms for recovering SFM[2]. We assume that the internal parameters of the cameras are known, and we seek to infer the shape of the scene, represented by a description of the surface of each object in some Euclidean reference frame, as well as the relative pose of the cameras. Since we cannot rely on individual point correspondences, we set up a cost functional that aims at matching regions, and integrate the irradiance over the entire domain of each image. We then develop gradient-based algorithms to estimate both the shape of the scene and the relative pose of the cameras.

1.1 Relation to previous work

In [25], a method is proposed to solve the multi-frame shape reconstruction of a smooth shape with constant radiance as the joint region segmentation of a collection of calibrated images. The reconstruction was remarkably robust with respect to image noise and deviation from the assumed

[*] Work supported by NSF IIS-0208197/CCR-0133736, ONR N00014-02-1-0720, and AFOSR F49620-03-1-0095.

[1] There are exceptions, which we discuss in section 1.1.

[2] Techniques that address this type of scenes by checking the photo-consistency of each "voxel" are available, although most require precise knowledge of the position of the cameras (algorithms that exploit the occluding boundaries to estimate both shape and camera pose are also available; see section 1.1 for more details).

piecewise constant radiance assumption, but also particularly sensitive to (extrinsic) calibration errors, thereby requiring precise positioning of the cameras. We extend their results to allow the position and orientation of each camera to be unknown, and therefore part of the inference process. Therefore, we estimate simultaneously surface shape, constant radiance, and camera motion. Our work is also closely related to the variational approach to stereo reconstruction, championed by Faugeras and Keriven [3]. They also assume the camera positions to be known, and therefore our work can be interpreted as a special case of of [3], extended to allow arbitrary camera pose. It should be note that our approach, like [3, 25, 16], is based on gradient descent algorithms, and therefore we always assume that the rough positioning of the cameras is available to be used as initial conditions to the algorithms. Therefore, our algorithm can be interpreted as a "refinement step" of the results of any multi-frame stereo calibration or structure from motion algorithm. However, the initialization needs not be precise. In section 4 we will show results obtained by initializing the camera positions and orientation by hand. Similarly, this work relates to shape carving techniques [5], since the reconstruction is achieved by evolving a volume in a way that is consistent with the image data. We explicitly enforce smoothness constraints, and therefore our approach does not work for arbitrary objects. However, for objects that satisfy the assumptions, our approach exhibits significant robustness to measurement noise. Furthermore, to the best of our knowledge, motion estimation has not been addressed within the context of shape carving, where the cameras are assumed to be calibrated. Since we assume constant or smooth radiance, most of the shape information concentrates at the occluding boundaries, and therefore our work relates to the literature on shape from silhouettes [2]. That work has indeed been extended to allow inference of camera motion as well as scene shape [1], although that was done within the framework of epipolar geometry. We estimate motion directly using a gradient procedure, and therefore do not require establishing correspondence between (real or virtual) points. Nevertheless, it should be noticed that the conditions for unique reconstruction are, of course, the same, and therefore we are subject to the same constraints as in silhouette-based methods. For instance, a unique camera motion cannot be recovered in the presence of symmetries of the object. Nevertheless, the shape can still be recovered, albeit relative to an unknown reference frame.

For the computational methods it uses, this paper is also related to a wealth of contributions in the field of region-based segmentation, starting from Mumford and Shah's pioneering work [6], and including [11, 24, 14, 15, 18, 20, 21, 22, 23, 13]. Our numerical implementation is based on Osher and Sethian's level set methods [8].

1.2 Outline and contributions of this paper

In Section 2, we will review the model proposed in [25] for joint image segmentation and shape reconstruction for a calibrated stereo rig. In section 3 we will extend this model to estimate motion parameters. Although this extension is conceptually straightforward, in practice its implementation is entirely non trivial; we report the calculations in section 3, which constitutes the original contribution of this paper, and discuss the implications in section 3.1. The resulting algorithms are tested on real and synthetic image sequences in section 4.

2 Reconstruction for calibrated cameras (review)

In [25], a model for joint image segmentation and shape reconstruction has been proposed. It is assumed that the scene is composed of a number of smooth, closed surfaces supporting smooth Lambertian radiance functions (or dense textures with spatially smooth statistics) and the background, which occupies the rest of the image. Under these assumptions, a subset of brightness (or texture) discontinuities correspond to occluding boundaries. These assumptions make the image segmentation problem well-posed, although not general.

2.1 Notation

Let S to be a smooth surface in $\mathbb{R}^3$ with local coordinates $(u, v) \in \mathbb{R}^2$. Let dA be its Euclidean area element, i.e. $dA = \|S_u \times S_v\|$; $\mathbf{X} = [X, Y, Z]^T$ the coordinates of a generic point on S. We measure n images, $I_i, i = 1, 2, \ldots, n$ and are given the internal calibration parameters, so that the camera is modeled as an ideal perspective projection: $\pi_i : \mathbb{R}^3 \rightarrow \Omega_i; \mathbf{X} \mapsto \mathbf{x}_i$, where $\mathbf{x}_i = [x_i, y_i]^T = [X_i/Z_i, Y_i/Z_i]^T$, $\Omega_i \subset \mathbb{R}^2$ is the domain of the image I_i, with area element $d\Omega_i$. We will use $\mathbf{X}_i = [X_i, Y_i, Z_i]^T$ to represent $\mathbf{X}$ in the "camera coordinates" with respect to the i-th camera. $\mathbf{X}$ and $\mathbf{X}_i$ are related by a rigid body transformation, described by an element of the Euclidean group $g_i \in SE(3)$, represented in coordinate by a rotation matrix $R_i \in SO(3)$ and a translation vector $T_i \in \mathbb{R}^3$, so that[3] $\mathbf{X}_i = g_i\mathbf{X} = R_i\mathbf{X} + T_i$. We describe the background B as a sphere with angular coordinates $\Theta = (\theta, \eta) \in \mathbb{R}^2$ that may be related in a one-to-one manner with the coordinates $\mathbf{x}_i$ of each image domain Ω_i through the mapping Θ_i. We assume that the background supports a radiance function $\mathbf{h} : \Theta \rightarrow \mathbb{R}_+$ and the surface supports another radiance function $\mathbf{f} : S \rightarrow \mathbb{R}_+$. We

[3] The reader will pardon an abuse of notation, since we mix the motion g_i and its representation (R_i, T_i); this is done for convenience of notation.

define the region $\tilde{\Omega}_i = \pi_i(S) \subset \Omega_i$ and denote its complement by $\tilde{\Omega}_i^c$. Although the perspective projection π_i is not one-to-one (and therefore not invertible), the operation of back-projecting a point $\mathbf{x}_i$ from $\tilde{\Omega}_i$ onto the surface S can be defined by tracing the ray starting from the i-th camera center and passing through $\mathbf{x}_i$, and defining the first intersection point as the back-projection of $\mathbf{x}_i$ onto S. Therefore, with an abuse of notation we denote this back-projection by $\pi_i^{-1} : \tilde{\Omega}_i \to S; \mathbf{x}_i \mapsto \mathbf{X}$.

In order to infer the shape of a surface S, one can impose a cost on the discrepancy between the projection of a model surface and the actual measurements. Such a cost, E, depends upon the surface S as well as upon the radiance of the surface $\mathbf{f}$ and of the background $\mathbf{h}$, as well as the motion g_i (through the projection π_i): $E = E(\mathbf{f}, \mathbf{h}, S, g_1, \dots, g_n)$. For simplicity, we indicate by g the collection of camera motions $g_1, \dots, g_n$. One can then adjust the shape of the model surface and radiances to match the measured images. Since the unknowns (surface S and radiances $\mathbf{f}, \mathbf{h}$) live in an infinite-dimensional space, we need to impose regularization. Therefore, the cost functional is a weighted average of three terms: $E(\mathbf{f}, \mathbf{h}, S) = E_{data}(\mathbf{f}, \mathbf{h}, S) + \alpha E_{geom}(S) + \beta E_{smooth}(\mathbf{f}, \mathbf{h}, S)$ where $\alpha, \beta \in \mathbb{R}^+$. The data fidelity term $E_{data}(\mathbf{f}, \mathbf{h}, S, g)$ quantifies the discrepancy between measured images and the images predicted by the model. For simplicity, we compute it in the sense of $\mathcal{L}^2$ on the image domain by $E_{data} = \sum_{i=1}^n \int_{\tilde{\Omega}_i} \big(\mathbf{f}(\pi_i^{-1}(\mathbf{x}_i)) - I_i(\mathbf{x}_i)\big)^2 d\mathbf{x}_i + \sum_{i=1}^n \int_{\tilde{\Omega}_i^c} \big(\mathbf{h}(\Theta_i(\mathbf{x}_i)) - I_i(\mathbf{x}_i)\big)^2 d\mathbf{x}_i$. $E_{smooth}(\mathbf{f}, \mathbf{h}, S)$ and $E_{geom}(S)$ measure the smoothness of the radiance functions and the surface respectively. They are given by $E_{geom} = \int_S dA = \text{area}(S)$, and $E_{smooth} = \int_S \|\nabla_S \mathbf{f}\|^2 dA + \int_B \|\nabla \mathbf{h}\|^2 d\Theta$, where ∇_S denotes the intrinsic gradient on the manifold S. (The exact definition and details on its computation can be found in [10]).

2.2 Computation of the gradient flow (review)

The data fidelity term may be measured To facilitate the computation of the variation with respect to S, we express these integrals over the surface S. This can be done using the characteristic functions $\chi_i(\mathbf{X}) = 1$ if $\mathbf{X}$ visible from the i-th camera and $\chi_i(\mathbf{X}) = 0$ otherwise. The data term E_{data} is therefore given by:

$$\sum_{i=1}^n \int_{\Omega_i} \rho_i^2(\mathbf{x}_i)\, d\mathbf{x}_i + \int_{\tilde{\Omega}_i} \Big(\big(\mathbf{f}(\pi_i^{-1}(\mathbf{x}_i)) - I_i(\mathbf{x}_i)\big)^2 - \rho_i^2(\mathbf{x}_i)\Big)\, d\mathbf{x}_i$$
$$= \sum_{i=1}^n \int_{\Omega_i} \rho_i^2(\mathbf{x}_i)\, d\mathbf{x}_i + \int_S \chi_i(\mathbf{X})(\tilde{\rho}_i^2(\mathbf{X}) - \rho_i^2(\pi_i(\mathbf{X})))\sigma_i(\mathbf{X}, N)\, dA$$

where $\tilde{\rho}_i(\mathbf{X}) = \mathbf{f}(\mathbf{X}) - I_i(\pi_i(\mathbf{X}))$ and $\rho_i(\mathbf{x}_i) = \mathbf{h}(\Theta_i(\mathbf{x}_i)) - I_i(\mathbf{x}_i)$. We use the fact that $\tilde{\Omega}_i$ is the projection of S in the i-th image and that the area measure $d\mathbf{x}_i$ of the image is related to the area measure dA of the surface by $d\mathbf{x}_i = (\mathbf{X}_i \cdot N_i)/Z_i^3 dA$, where N is the inward unit normal to S, and N_i is N with respect to the coordinates of the i-th camera. $\sigma_i(\mathbf{X}, N)$ is a shorthand notation for $(\mathbf{X}_i \cdot N_i)/Z_i^3$. In the simpler case where both radiance functions $\mathbf{f}$ and $\mathbf{h}$ are constant, the overall cost functional can be simplified to:

$$E_{constant} = \alpha \int_S dA + \sum_{i=1}^n \int_{\Omega_i} \rho_i^2(\mathbf{x}_i)\, d\mathbf{x}_i$$
$$+ \sum_{i=1}^n \int_S \chi_i(\mathbf{X})(\tilde{\rho}_i^2(\mathbf{X}) - \rho_i^2(\pi_i(\mathbf{X})))\sigma_i(\mathbf{X}, N) dA.$$

This simplification relates to the approach of Chan and Vese [24] who considered a piece-wise constant version of the Mumford-Shah functional for 2-D images in the level set framework [17, 19].

The gradient flow of the cost functional E has been derived in [25]. The flow corresponding to the data fidelity term is given by

$$\frac{dS}{dt} = \frac{1}{z^3}\Big((\mathbf{f}-\mathbf{h})\big[(I-\mathbf{f})+(I-\mathbf{h})\big](\nabla\chi \cdot S) + 2\chi(I-\mathbf{f})(\nabla\mathbf{f} \cdot S)\Big) N \tag{1}$$

Notice that this flow depends only upon the image values, *not the image gradient*, which makes it more robust to image noise when compared to other variational approaches to stereo (i.e. less prone to become "trapped" in local minima).

The gradient flow corresponding to the smoothness term, also derived in [25], is given by

$$\frac{dS}{dt} = \big(\mathrm{II}(\nabla_S \mathbf{f} \times N) - \|\nabla_S \mathbf{f}\|^2 H\big) N \tag{2}$$

where the second fundamental form of $\nabla_S \mathbf{f} \times N$ is computed as

$$\mathrm{II}(\nabla_S \mathbf{f} \times N) = \frac{\mathbf{f}_u^2 g - 2\mathbf{f}_u \mathbf{f}_v f + \mathbf{f}_v^2 e}{EG - F^2}$$

and the coefficients e, f, g of the second fundamental form are given by $e = \langle N, S_{uu} \rangle$, $f = \langle N, S_{uv} \rangle$, and $g = \langle N, S_{vv} \rangle$ and the coefficients of the first fundamental form are $E = S_u \cdot S_u$, $F = S_u \cdot S_v$, $G = S_v \cdot S_v$.

The term $\nabla\chi \cdot S$ must be defined in the distributional sense because the characteristic function χ is discontinuous. It can be shown that

$$\nabla\chi \cdot S = -\kappa_u \|S\|^2 \delta(S \cdot N) \tag{3}$$

where κ_u denotes the normal curvature of S in the u-direction (the direction S where $S \cdot N = 0$). The overall flow can be computed by summing the flow corresponding to each component of the cost functional. For the case of constant radiance, for instance, one gets

$$\begin{aligned} \frac{dS}{dt} &= \frac{1}{z^3}(\mathbf{f}-\mathbf{h})\big[(I-\mathbf{f})+(I-\mathbf{h})\big](\nabla\chi\cdot S)N \\ &= -\frac{\kappa_u\|S\|^2}{z^3}(\mathbf{f}-\mathbf{h})\big[(I-\mathbf{f})+(I-\mathbf{h})\big]\delta(S\cdot N)N. \end{aligned}$$

3 Evolving the motion parameters

We now consider the same energy functional E as a function of the motion parameters $g_i \in SE(3)$. We will use the exponential parameterization of $SE(3)$ via the twist coordinates $\xi_i \in \mathbb{R}^6$. The parameterization is established by a map from $\mathbb{R}^6$ to the Lie algebra $se(3)$ via $\xi_i \mapsto \widehat{\xi_i}$, which is exponentiated to lead the motion $g_i = \exp(\widehat{\xi_i}) \in SE(3)$. The reader can consult [7] for more details on twists and exponential coordinates for rigid motions.

What matters, however, is that we can represent locally g with a six-parameter vector ξ. We will denote the local parameterization via $g_i = g_i(\xi)$ where $\xi = (\xi_{i1}, \ldots, \xi_{i6})$ for each camera image I_i. Notice that the only term in our energy functional E which depends upon ξ_i is the corresponding fidelity term in E_{data} (due to the dependence of π_i^{-1} and Θ_i on ξ_i): $E_{data,i}(S,\mathbf{f},\mathbf{h},\xi_i)$ is therefore given by

$$\int_{\tilde{\Omega}_i}\big(\mathbf{f}(\pi_i^{-1}(\bar{\mathbf{x}}))-I_i(\bar{\mathbf{x}})\big)^2 d\mathbf{x}_i+\int_{\tilde{\Omega}_i^c}\big(\mathbf{h}(\Theta_i(\bar{\mathbf{x}}))-I_i(\bar{\mathbf{x}})\big)^2 d\mathbf{x}_i \tag{4}$$

If we let $\bar{c}_i = \partial\tilde{\Omega}_i$ denote the boundary of $\tilde{\Omega}_i$ then we may express the partial derivative of E with respect to one of the calibration parameters ξ_{ij}. $\frac{\partial E}{\partial \xi_{ij}}$ is given by the sum of three terms: a boundary term, a foreground term and a background term, given respectively by

$$\begin{aligned} &\int_{\bar{c}_i}\Big(\big(\mathbf{f}(\pi_i^{-1}(\bar{\mathbf{x}}))-I_i(\bar{\mathbf{x}})\big)^2-\big(\mathbf{h}(\Theta_i(\bar{\mathbf{x}}))-I_i(\bar{\mathbf{x}})\big)^2\Big)\langle\frac{\partial\,\bar{c}_i}{\partial\xi_{ij}},\bar{n}_i\rangle\, d\bar{s} \\ &\quad 2\int_{\tilde{\Omega}_i}\big(\mathbf{f}(\pi_i^{-1}(\bar{\mathbf{x}}))-I_i(\bar{\mathbf{x}})\big)\Big\langle\nabla_S\mathbf{f}\big(\pi_i^{-1}(\bar{\mathbf{x}})\big),\frac{\partial}{\partial\xi_{ij}}\pi_i^{-1}(\bar{\mathbf{x}})\Big\rangle d\mathbf{x}_i \\ &\qquad 2\int_{\tilde{\Omega}_i^c}\big(\mathbf{h}(\Theta_i(\bar{\mathbf{x}}))-I_i(\bar{\mathbf{x}})\big)\Big\langle\nabla_B\mathbf{h}\big(\Theta_i(\bar{\mathbf{x}})\big),\frac{\partial}{\partial\xi_{ij}}\Theta_i(\bar{\mathbf{x}})\Big\rangle d\mathbf{x}_i \end{aligned}$$

In the boundary term, $d\bar{s}$ denotes the arc-length measure of $\bar{c}_i$, and $\bar{n}_i$ denotes its outward unit normal. In the foreground term, ∇_S denotes the natural gradient operator on the surface S, while in the background term, ∇_B denotes the gradient operator with respect to the angular coordinates of the background B.

It is convenient to express the contour integral around $\bar{c}_i(\bar{s})$ in the image plane as a contour integral around $C_i(s)$ on the surface S instead, (where $\pi_i(C_i)=\bar{c}_i$ and where s is the arc-length parameter of C_i). They are related by

$$\langle\frac{\partial\,\bar{c}_i}{\partial\xi_{ij}},\bar{n}_i\rangle\, d\bar{s} = \Big\langle\frac{\partial}{\partial\xi_{ij}}\pi_i(C_i),\frac{\partial}{\partial s}J\pi_i(C_i)\Big\rangle\, ds,$$

where $J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and, therefore, the expression above is given by

$$\begin{aligned} &\frac{1}{z_i^3}\left\langle\frac{\partial\mathbf{x}_i}{\partial s},\begin{bmatrix} 0 & -z_i & y_i \\ z_i & 0 & -x_i \\ -y_i & x_i & 0 \end{bmatrix}\frac{\partial\mathbf{x}_i}{\partial\xi_{ij}}\right\rangle ds \\ &= \frac{1}{z_i^3}\Big\langle\frac{\partial\mathbf{x}_i}{\partial s},\frac{\partial\mathbf{x}_i}{\partial\xi_{ij}}\times\mathbf{x}_i\Big\rangle ds = \frac{1}{z_i^3}\Big\langle\frac{\partial\mathbf{x}_i}{\partial\xi_{ij}},\mathbf{x}_i\times\frac{\partial\mathbf{x}_i}{\partial s}\Big\rangle ds \\ &= \frac{\|\mathbf{x}_i\|}{z_i^3}\Big\langle\frac{\partial\mathbf{x}_i}{\partial\xi_{ij}},N_i\Big\rangle ds \end{aligned}$$

since $\mathbf{x}_i$ and $\frac{\partial\mathbf{x}_i}{\partial s}$ are perpendicular tangent vectors to S. Thus, the boundary term written as an integral on the surface S (along the occluding contour C_i) has the following form.

$$\int_{C_i}\Big(\big(\mathbf{f}-I_i\big)^2-\big(\mathbf{h}-I_i\big)^2\Big)\frac{\|\mathbf{x}_i\|}{z_i^3}\Big\langle\frac{\partial g_i}{\partial\xi_{ij}},N_i\Big\rangle ds \tag{5}$$

The first step in rewriting the foreground/background integrals is to re-express the derivative of the back-projected 3D point $\mathbf{x} = \pi_i^{-1}(\bar{\mathbf{x}}, g_i)$ with respect to the calibration parameter ξ_{ij} in terms of the derivative of the forward projection $\pi_i(\mathbf{x}, g_i) = \pi(g_i(\mathbf{x}, g_i))$, since π_i has an analytic form while π_i^{-1} does not. We begin by fixing a 2D image point $\bar{\mathbf{x}}$ and note that $\bar{\mathbf{x}} = \pi_i\big(\mathbf{x}(\bar{\mathbf{x}}, g_i), g_i\big)$ where $\mathbf{x}(\bar{\mathbf{x}}, g_i) = \pi_i^{-1}(\bar{\mathbf{x}}, g_i) = g_i^{-1}\big(\pi^{-1}(\bar{\mathbf{x}}), g_i\big)$ and thus differentiation with respect to ξ_{ij} yields:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\xi_{ij}}\pi_i\big(\mathbf{x}, g_i\big) = \frac{\partial\pi_i}{\partial\mathbf{x}}\frac{\partial\mathbf{x}}{\partial\xi_{ij}}+\frac{\partial\pi_i}{\partial\xi_{ij}} \\ &= \frac{1}{z_i^2}\begin{bmatrix} z_i & 0 & -x_i \\ 0 & z_i & -y_i \end{bmatrix}\frac{\partial g_i}{\partial\mathbf{x}}\frac{\partial\mathbf{x}}{\partial\xi_{ij}}+\frac{1}{z_i^2}\begin{bmatrix} z_i & 0 & -x_i \\ 0 & z_i & -y_i \end{bmatrix}\frac{\partial g_i}{\partial\xi_{ij}} \\ &\begin{bmatrix} z_i & 0 & -x_i \\ 0 & z_i & -y_i \end{bmatrix}\frac{\partial g_i}{\partial\mathbf{x}}\frac{\partial\mathbf{x}}{\partial\xi_{ij}} = -\begin{bmatrix} z_i & 0 & -x_i \\ 0 & z_i & -y_i \end{bmatrix}\frac{\partial g_i}{\partial\xi_{ij}} \end{aligned} \tag{6}$$

Notice, though, that (6) does not uniquely specify $\partial\mathbf{x}/\partial\xi_{ij}$ but merely gives a necessary condition. We must supplement (6) with the additional constraint that $\partial\mathbf{x}/\partial\xi_{ij}$ must be orthogonal to the unit normal N of S at the point $\mathbf{x}$ in order to obtain a unique solution.

$$\frac{\partial\mathbf{x}}{\partial\xi_{ij}}\cdot N = 0 \tag{7}$$

Now, combining equations (6) and (7), we have

$$\begin{aligned} &\begin{bmatrix} z_i & 0 & -x_i \\ 0 & z_i & -y_i \\ N_{ix} & N_{iy} & N_{iz} \end{bmatrix}\frac{\partial g_i}{\partial\mathbf{x}}\frac{\partial\mathbf{x}}{\partial\xi_{ij}} = -\begin{bmatrix} z_i & 0 & -x_i \\ 0 & z_i & -y_i \\ 0 & 0 & 0 \end{bmatrix}\frac{\partial g_i}{\partial\xi_{ij}} \\ &\frac{\partial\mathbf{x}}{\partial\xi_{ij}} = -\left(\frac{\partial g_i}{\partial\mathbf{x}}\right)^{-1}\left(\mathcal{I}-\frac{\mathbf{x}_i\otimes N_i}{\mathbf{x}_i\cdot N_i}\right)\frac{\partial g_i}{\partial\xi_{ij}} \end{aligned} \tag{8}$$

The second step proceeds in the same manner as outlined earlier in rewriting the data fidelity terms in E_{data} by noting that the measure in the image domain $d\mathbf{x}_i$ and the area measure on the surface dA are related by $d\mathbf{x}_i = \sigma(\mathbf{x}_i, N_i)\, dA$ where $\sigma(\mathbf{x}_i, N_i) = (\mathbf{x}_i \cdot N_i)/z_i^3$.

$$
\begin{aligned}
&2\int_{\tilde{\Omega}_i} (\mathbf{f} - I_i) \left\langle \nabla_S \mathbf{f}(\pi_i^{-1}(\bar{\mathbf{x}})), \frac{\partial}{\partial \xi_{ij}} \pi_i^{-1}(\bar{\mathbf{x}}) \right\rangle d\mathbf{x}_i \\
&= 2\int_{\pi_i^{-1}(\tilde{\Omega}_i)} (\mathbf{f} - I_i) \left\langle \nabla_S \mathbf{f}(\mathbf{x}), \frac{\partial \mathbf{x}}{\partial \xi_{ij}} \right\rangle \frac{\mathbf{x}_i \cdot N_i}{z_i^3}\, dA
\end{aligned}
$$

The integrand above can be written more explicitly as $-\frac{(\mathbf{f}-I_i)}{z_i^3} \left\langle \nabla_S \mathbf{f}(\mathbf{x}), \left(\frac{\partial g_i}{\partial \mathbf{x}}\right)^{-1} \left((\mathbf{x}_i \cdot N_i) \frac{\partial g_i}{\partial \xi_{ij}} - \left(\frac{\partial g_i}{\partial \xi_{ij}} \cdot N_i\right) \mathbf{x}_i\right) \right\rangle$

A similar derivation can be followed for the background term. The calculations above yield the gradient of the cost functional E with respect to the local coordinates of the motion parameters ξ, $\frac{\partial E}{\partial \xi}$. This is transformed into a vector in the tangent space to the motion parameters g via the lifting to the Lie algebra, $\widehat{(\frac{\partial E}{\partial \xi})} \in se(3)$. The evolution of the motion parameters is finally given by The final expression for the flow with respect to the local coordinates of the motion parameters is given by

$$
\frac{dg}{dt} = \widehat{\left(\frac{\partial E}{\partial \xi}\right)} g dt \tag{9}
$$

To complete the algorithm, one or more steps of the flow (9) are alternated to one or more steps of the flow (1), until the value of the cost functional reaches steady state (it is easy to prove that, with an appropriate choice of step-size, every step lowers the value of the cost functional).

3.1 Uniqueness, or lack thereof

Note that the flow converging to steady-state guarantees that the shape and motion parameters converge *somewhere*, but in general it does not guarantee that they converge to the correct shape or relative pose of the camera. For instance, consider the case of a sphere, imaged by a number of cameras distributed around a circumference centered at the center of the sphere. The image of the scene in each camera is identical, and therefore there is no way to tell where the cameras are. Nevertheless, one can conclude from the images that the scene is a sphere (assuming the cameras are in general position), and minimize the discrepancy of the model image (the projection of the estimated sphere) from the measured images.

More in general, Euclidean symmetries in shape will generate ambiguities in the estimates of relative pose. One can have continuous symmetries (such as in the example of the sphere) or discrete symmetries (such as in the case of a homogeneous cube).

Nevertheless, if one is interested in the shape of the scene, regardless of the positioning of the camera, the alternation of the flow (9) and (1) will indeed provide an estimate of shape that simultaneously explains each given image.

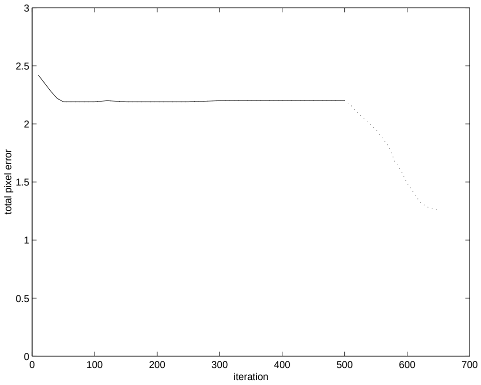

Figure 2: *Value of total squared reprojection error (E_{data} in our cost functional) initially as only the shape is evolved while fixing the camera poses given by an external calibration procedure (solid line, 500 steps) and subsequently as the camera poses are also evolved (dotted line, 150 steps). Units are the sum over all pixels in each image of the squared intensity differences between the pixel and model intensities (units are 10^9).*

In general, a full-fledged analysis of the uniqueness of the minimizers of the functional we describe is well beyond the scope of this paper. However, some conclusions may be drawn from the analysis of SFM for the case of point features, for which we refer the reader to [4].

Naturally, since the algorithm we propose is a gradient flow, convergence is only guaranteed locally, since the algorithm can get trapped in local minima. However, in every experiment we have performed, some of which are reported in the next section, we have seldom experienced convergence to local minima despite coarse initialization.

4 Experiments

In figure 1 we show a few images of a test scene meant to challenge the assumptions common to most SFM algorithms. Our scheme is design to work under these assumptions.

In figure 3 we show the evolution of the estimate of shape if the pose of the cameras is taken to be the result of an external calibration procedure, and the significant improvement that follows when the camera pose is allowed to vary and is part of the inference process. This improvement is quantified in figure 2.

In figure 4 we show the reprojection error, i.e. the best estimate of shape projected onto the image according to the best estimate of the camera pose, for when the camera parameters are fixed (top) or allowed to vary (middle). To emphasize the improvement that follows the evolution of the motion parameters, we also show the reprojection error when the true shape, but wrong parameters, are used.

Finally, to emphasize the importance of incorporating

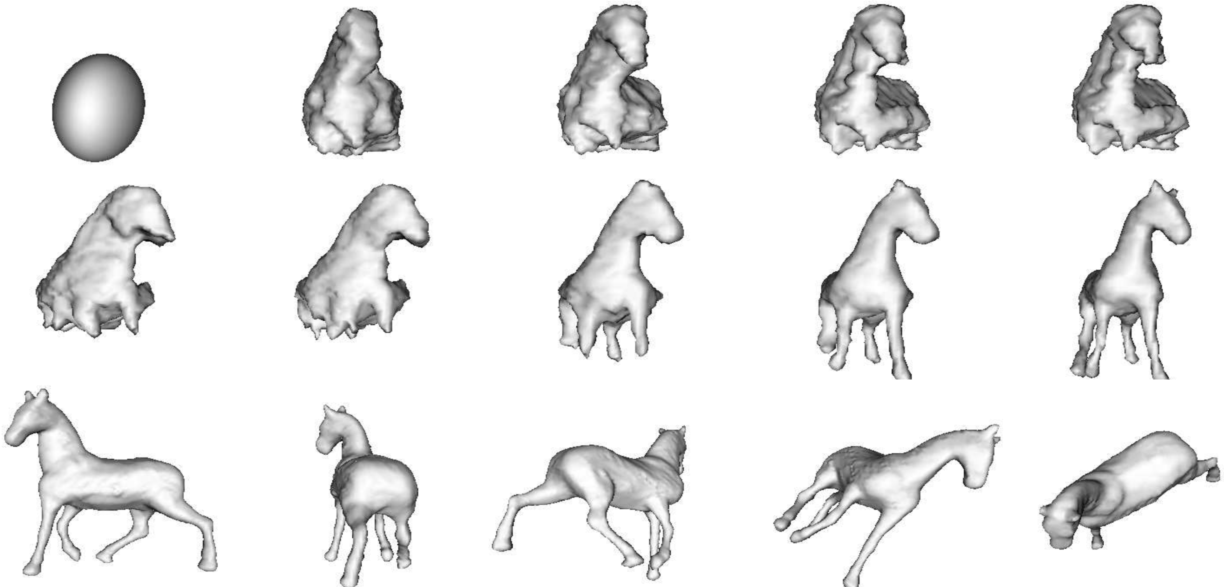

Figure 3: *Evolution of the estimate of shape when the camera pose is fixed with an external calibration procedure (top, after 0, 50, 100, 300 and 500 steps); evolution of estimate of shape joined with the estimate of the motion parameters (middle, after 30, 60, 90, 110 and 150 steps); final estimate from several viewpoints (bottom).*

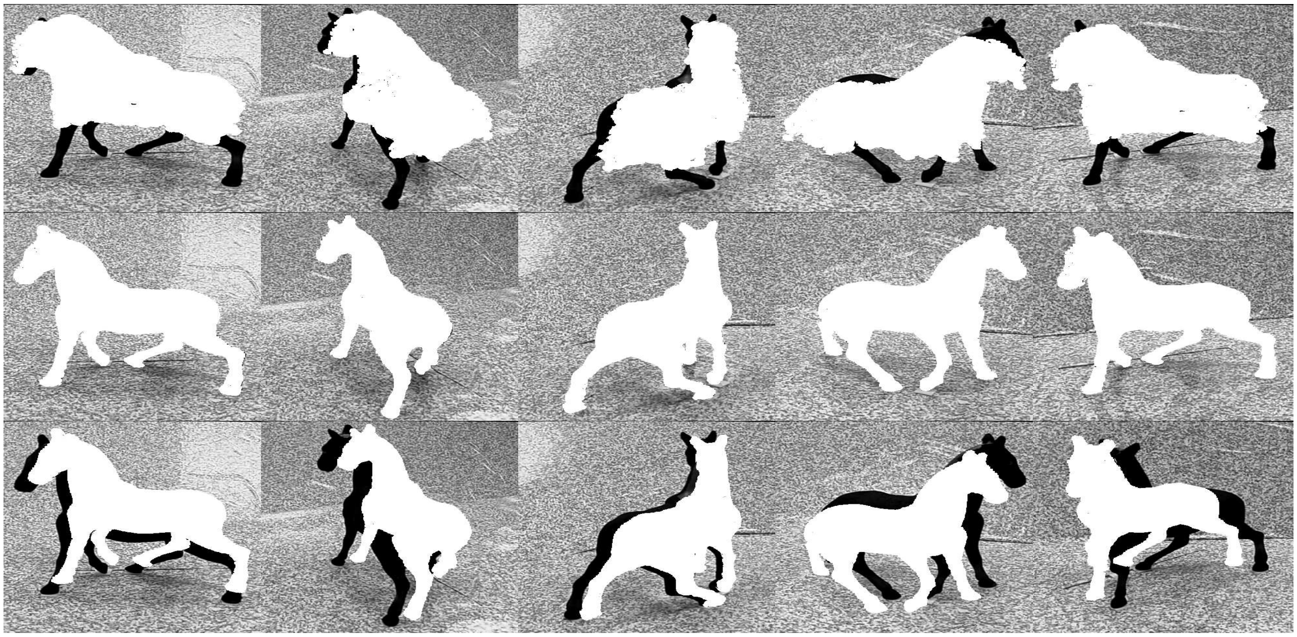

Figure 4: *Reprojection error when the camera pose is fixed with an external calibration procedure (top), and when camera pose is estimated along with scene shape (middle). Reprojection error for the correct shape if the camera parameters were fixed (bottom).*

the segmentation procedure behind our model as **during** the process of shape reconstruction and pose estimation rather than as a "first step", we show in figure 4 the results of constructing the visual hull directly from the segmented images. Such a procedure, which relates much more directly to space carving and shape-from-silhouettes than our approach may seem quite tempting since the images are individually easy to segment. However, as can be seen in the figure, the final reconstruction obtained using this serial method of "first segment then reconstruct" suffers terribly in the presence of calibration errors. To highlight this point we also show the visual hull reconstruction using the final updated camera poses obtained after the evolution illustrated in the previous figures.

5 Conclusions

We have presented an algorithm to estimate the shape of a scene composed of smooth surfaces with constant radiance as well as the relative pose of a collection of cameras. We define a cost functional that penalizes the discrepancy between the measured images and the projection of the estimated model onto the image, as well as regularizing terms to enforce the smoothness assumptions. We define a gradient flow procedure that is guaranteed to minimize (locally) the cost functional. As the experiments show, our algorithm is very robust to image noise and to the initialization of the scene shape. It does require initialization of the relative pose of the cameras, although a manually inputed guess is usually sufficient. It performs well under the assumptions it is designed for. It does not work when the scene has non-smooth radiance, a condition that allows other algorithms for SFM to work well.

References

[1] K. Astrom, R. Cipolla, and P. J. Giblin. Motion from the frontier of curved surfaces. pages 269–275, 1995.

[2] R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *Int. J. of Computer Vision, 9 (2)*, 1992.

[3] O. D. Faugeras and R. Keriven. Variational principles, surface evolution pdes, level set methods and the stereo problem. *INRIA Technical report*, 3021:1–37, 1996.

[4] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision.* Cambridge University Press, 2000.

[5] K. Kutulakos and S. Seitz. A theory of shape by space carving. In *Proc. of the Intl. Conf. on Comp. Vision*, 1998.

[6] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. on Pure and Applied Mathematics*, 42:577–685, 1989.

[7] R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[8] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi equations. *J. of Comp. Physics*, 79:12–49, 1988.

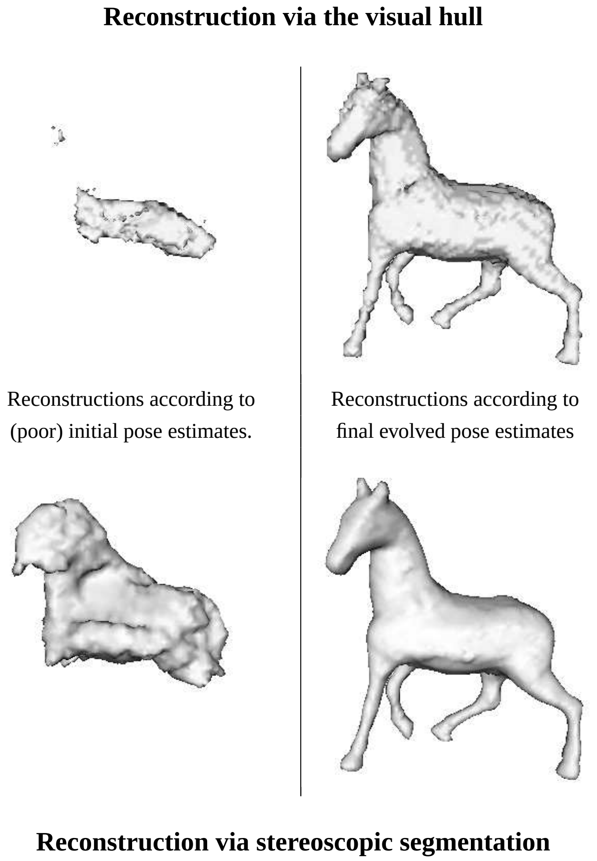
Reconstruction via the visual hull
Reconstructions according to (poor) initial pose estimates.
Reconstructions according to final evolved pose estimates
Reconstruction via stereoscopic segmentation

Figure 5: *Direct shape reconstructions using the visual hull of the (pre)segmented images are shown on top both for the initial incorrect pose estimates (left) and for the final improved pose estimates recovered by our algorithm (right). Our shape reconstructions in both cases are shown below. Note that while the photo hull reconstruction on the top-right gives a decent shape, such a reconstruction was only possible by using the final pose estimates yielded simultaneously with the bottom-right shape by our algorithm. Also note that the initial shape (bottom-left) recovered by our algoritm* **before** *allowing it to evolve the pose estimates in order to obtain the final shape (bottom-right) still gives at least a somewhat recognizable coarse scale representation of the horse despite the significant calibration errors whereas the corresponding direct reconstruction of the visual hull (top-left) is not at all recognizable.*

[9] R. Tsai. A versatile camera calibration technique for high accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE J. Robotics Automat.*, RA-3(4):323–344, 1987.

[10] M. Bertalmio, L. Cheng, S. Osher and G. Sapiro. Variational Problems and Partial Differential Equations on Implicit Surfaces. *Journal of Computational Physics*, pp 759–780, vol 174, No. 2, Dec 2001.

[11] V. Caselles, R. Kimmel, G. Sapiro. Geodesic Active Contours. *Int. J. Computer Vision*, vol. 22, no. 1, pp 61–79, Feb. 1997.

[12] A. Chakraborty and J. Duncan. Game-Theoretic Integration for Image Segmentation. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 1, pp. 12–30, Jan. 1999.

[13] A. Chakraborty, L. Staib, and J. Duncan. Deformable Boundary Finding in Medical Images by Integrating Gradient and Region Information. *IEEE Trans. Medical Imaging*, vol. 15, no. 6, pp. 859–870, Dec. 1996.

[14] L. Cohen. On Active Contour Models and Balloons. *CVGIP: Image Understanding*, vol. 53, pp. 211–218, 1991.

[15] M. Kass, A. Witkin, and D. Terzopoulos, Snakes: Active Contour Models. *Int. Journal of Computer Vision*, vol. 1, pp. 321–331, 1987.

[16] R. Kimmel. 3D shape reconstruction from autostereograms and stereo. *Journal of Visual Communication and Image Representation*, 13:324-333, March 2002.

[17] S. Osher and C.-W. Shu. High-order Essentially Nonoscillatory Schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 28(4):907–922, 1991.

[18] R. Ronfard, Region-Based Strategies for Active Contour Models. *Int. J. Computer Vision*, vol. 13, no. 2, pp. 229–251, 1994.

[19] Panagiotis E. Souganidis, Approximation Schemes for Viscosity Solutions of Hamilton-Jacobi Equations. *J. Differential Equations*, vol. 59, no. 1, pp 1–43, 1985.

[20] H. Tek and B. Kimia. Image Segmentation by Reaction Diffusion Bubbles. *Proc. Int. Conf. Computer Vision*, pp. 156–162, 1995.

[21] D. Terzopoulos and A. Witkin. Constraints on Deformable Models: Recovering Shape and Non-rigid Motion. *Artificial Intelligence*, vol. 36, pp. 91–123, 1988.

[22] S. Zhu, T. Lee, A. Yuille. Region Competition: Unifying snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation. *Proc. Int. Conf. Computer Vision*, pp. 416–423, 1995.

[23] S. Zhu and A. Yuille. Region Competition: Unifying snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 884–900, Sep. 1996.

[24] T. Chan and L. Vese. Active contours without edges *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, Feb. 2001.

[25] A. Yezzi and S. Soatto. Stereoscopic segmentation. In *Proc. of the Intl. Conf. on Computer Vision*, pages 59–66, 2001.