Variational Multiframe Stereo in the Presence of Specular Reflections

Hailin Jin[†] Anthony J. Yezzi[‡] Stefano Soatto[§][*]

[†] Electrical Engineering, Washington University, Saint Louis, MO 63130
[‡] Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332
[§] Computer Science, UCLA, Los Angeles, CA 90095

Abstract

We consider the problem of estimating the surface of an object from a calibrated set of views under the assumption that the reflectance of the object is non-Lambertian. In particular, we consider the case when the object presents sharp specular reflections. We pose the problem within a variational framework and use fast numerical techniques to approach the local minimum of a regularized cost functional.

1 Introduction

Multiframe stereo consists of estimating the three-dimensional shape of an object from a collection of views and is one of the classical problems of Computer Vision. In particular, we concentrate on the calibrated case, where both the internal parameters of each camera and their relative configuration is known[1]. The problem can then be decomposed, conceptually, into two steps: *correspondence* and *triangulation*. Correspondence refers to the problem of establishing which point corresponds to which in different images. This is typically addressed by making assumptions on the radiance distribution of the object as well as on the illumination from the environment. Once point correspondence is available, together with a statistical description of the correspondence error, triangulation can be posed as an optimization problem to determine the three-dimensional position of corresponding points in a common reference frame (see [4] for extensive references).

Of the two components, the first is by far the hardest since correspondence cannot be established for every point on the scene (the so-called "aperture" problem); furthermore, prior assumptions on the reflectance of the scene cannot be validated from the data, which makes it difficult to detect mismatches, outliers and other ambiguities in the correspondence.

Championed by Faugeras and Keriven [5], several methods have been proposed in recent years to address this problem by posing the multiframe stereo problem within a variational framework: a cost functional that measures the discrepancy between corresponding regions is minimized with respect to the shape of the object in space. This avoids having to specify point correspondences and allows imposing a certain degree of global regularity on the estimated shape. Discrepancy criteria used include normalized cross-correlation, total variation, $\mathcal{L}^1$ and $\mathcal{L}^2$ norms. These methods, however, do not solve all problems in multiframe stereo. Even though the cost functionals described measure discrepancy between corresponding regions, they still rely on the assumption that the appearance of a region does not

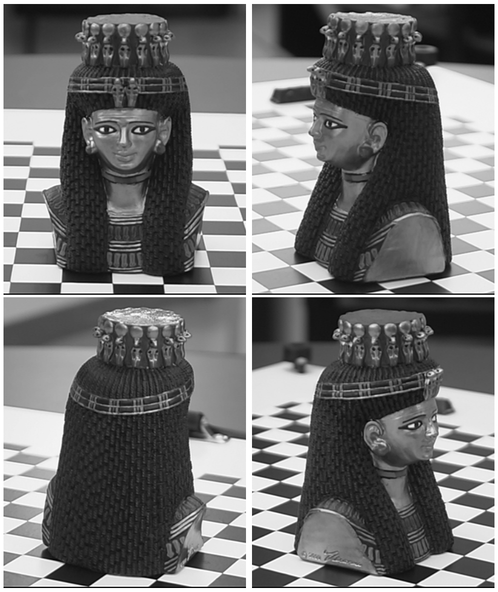

Figure 1: *Original dataset, consisting of 27 images of an object with specular surfaces (the face of the statue is lacquered in gold). This presents a challenge to traditional stereo algorithms, as shown in Figure 2 since the appearance changes dramatically depending on the position of the object relative to the light.*

[*] This research is supported in part by ARO grant DAAD19-99-1-0139 and Intel grant 8029.

[1] Several established methods are available to calibrate a stereo system and are described in textbooks, for instance in [4].

change significantly across different views[2].

This model is patently violated for objects that exhibit specular reflections. In Figure 1 we show a few views of a statue covered by a golden shiny finish; as it can be seen, the appearance changes significantly from different points of view since the reflection moves as the object moves relative to the light (notice for instance the dramatic changes in the appearance of the forehead, or of the top of the crown). Figure 2 shows the result of applying a variational stereo algorithm based on a cross-correlation discrepancy on the dataset of Figure 1. As it can be seen, details are lost, indeed even macroscopic structures such as the crown.

This paper addresses precisely this problem: how to perform multiframe stereo reconstruction within a variational framework in the presence of significant departures from a Lambertian model. In particular, we concentrate on surfaces with a concentrated specular component, so that sharp highlights are present in significantly different positions in different images.

In principle, the motion of specular reflections among images provides information about the shape of the object, under suitable assumptions on the illumination of the environment and on the reflectance distribution. However, like in the correspondence problem, such information relies on prior assumptions on the global illumination that cannot be validated from the data. From images alone it is not possible to reconstruct shape, reflectance and light distribution uniquely [10] (one can easily construct counter-examples of purely reflective objects of different shape that, once placed in environments radiating light with different distributions, produce exactly the same set of images). Therefore, we avoid building a universal imaging model and trying to invert it. Instead, the basic idea underlying this paper is to exploit the defining property of specular reflections (i.e. the fact that they move on the surface of the object when seen from different views) in order to identify the subset of three or more views of the same point that is unaffected by specularity.

1.1 Relation to previous work

This work naturally relates to several attempts to handle specular reflections in stereo matching and reconstruction. Bhat and Nayar [1, 2] consider the likelihood of correct stereo matching by analyzing the relationship between stereo vergence and surface roughness, and also propose a trinocular system where only two images are used at a time in the computation of depth at a point. Brelstaff and Blake [3] excise the specularities as a pre-processing step; similar techniques are used also by Okutomi and Kanade [8], while Nayar et al. [7] have considered using polarized filters to remove specularities. Ikeuchi [6] formulates the reconstruction problem for specular surface in a photometric stereo setting.

There is also a close relation between our work and that of Faugeras and Keriven [5], who cast the traditional multiframe stereo in a variational framework and use level set methods to solve it. They address the correspondence problem by best approximating the brightness constancy assumption at every point in the image[3], thus obtaining in effect a dense correspondence wherever the brightness gradient is non-zero.

This work also relates to the general problem of estimating reflectance properties as well as shape from sequences of images; for instance, Yu et al. use known shape to estimate global illumination [10].

We formulate a variational problem, derive the optimality conditions and design a partial differential equation (PDE) that converges to a solution that satisfies the necessary conditions. The PDE is solved numerically using level set methods, and therefore our work is related to the vast literature on level sets initiated by Osher and Sethian [9].

1.2 Contributions of this paper

We address the problem of multiframe stereo in the presence of specular reflections. To the best of our knowledge, we are the first to do so within a variational framework. We show that under suitable conditions (that each point being affected by a specularity is visible in at least three views) it is possible to estimate a suitable model of the shape of the scene (the smoothest shape that is photometrically consistent with the data). We test our algorithms on real images sequences, and we report some preliminary although very promising experimental results.

2 Variational Frameworks for Stereo

2.1 Notation and the Lambertian assumption

Let S be a smooth surface in $\mathbb{R}^3$ corresponding to an object in a 3D scene being imaged from multiple viewpoints. We indicate a point on this surface by $\mathbf{x}$, and the inward unit normal vector by $N(\mathbf{x})$. Most stereo algorithms, overtly or covertly, assume that the scene is *Lambertian*; that is its radiance $\mathbf{f}$ is independent of the viewing direction. Under this assumption, the radiance can be represented by a function $\mathbf{f} : S \to \mathbb{R}$ on the surface, which captures both the photometric properties of the surface as well as the illumination properties of the scene.

The radiance function $\mathbf{f}$ together with the surface S is sufficient to tell us how to construct any 2D image of a Lambertian surface given the location and orientation of the

[2] Normalized cross-correlation is invariant to absolute value and therefore accounts for invariance to contrast.

[3] This is done by looking for corresponding patches that maximize a normalized cross-correlation criterion, the underlying assumption being that of brightness constancy of corresponding points modulo local contrast and scale.

camera. One of the consequences of this simple model is that the reflected light intensity coming from a particular point $\mathbf{x} \in S$ is uniform in all directions, thereby giving rise to the same image intensities $I_i(m_i)$ and $I_j(m_j)$ (or *irradiances*) at the corresponding locations m_i and m_j in two different images. These corresponding locations m_i and m_j are given by ideal perspective projections $\pi_i : \mathbb{R}^3 \rightarrow \Omega_i$ and $\pi_j : \mathbb{R}^3 \rightarrow \Omega_j$ (where Ω_i and Ω_j denote the image planes of I_i and I_j respectively): $m_i = \pi_i(\mathbf{x})$ and $m_j = \pi_j(\mathbf{x})$.

Although the radiance function is necessary to generate a 2D image of a Lambertian surface, it is not necessary to test the hypothesis that a particular point $\mathbf{x} \in \mathbb{R}^3$ belongs to S. If we are given at least two images of the surface from different viewpoints, then $\mathbf{x} \in S$ implies that $I_i(\pi_i(\mathbf{x})) = \mathbf{f}(\mathbf{x})$ and $I_j(\pi_j(\mathbf{x})) = \mathbf{f}(\mathbf{x})$ which in turn implies that $I_i(\pi_i(\mathbf{x})) - I_j(\pi_j(\mathbf{x})) = 0$. This last expression, which does not involve the radiance, simply amounts to measuring the similarity between image intensities at points in two different images and is the basis of almost every stereo algorithm. Clearly, though, this intensity matching criterion rests strongly upon the Lambertian assumption and is no longer sensible when this assumption is violated.

2.2 Variational Lambertian models

We may use the intensity matching criterion discussed above for the Lambertian case in order to construct objective functions which we may use to compare the overall matching score induced by different choices for the surface S. This is the basic paradigm behind most variational approaches to stereo.

We start by recalling the variational level set framework proposed by Faugeras and Keriven for multi-frame stereo reconstruction. Their approach was to choose the surface S that minimized an energy functional in the form of a weighted area

$$E(S) = \int_S \Phi(\mathbf{x})\, dA, \tag{1}$$

where dA denotes the regular Euclidean area measure of the surface. Energy functionals of this sort are geometric in nature since dA is an intrinsic measure dictated only by the geometry of the surface S and is therefore invariant to different parameterizations of the same surface. The key to making this model useful for stereo reconstruction is to choose $\Phi(\mathbf{x})$ to be smaller at points which are close to desirable surface locations (in this case, locations which induce good stereo matching) and larger at points that are further away. Once Φ has been chosen, a surface S which minimizes (1), at least locally, can be obtained via gradient descent evolution by introducing an artificial time variable t. This is done by deforming some initial guess for S according to the following gradient flow

$$\frac{\partial S}{\partial t} = \Phi H N - \nabla \Phi \cdot N, \tag{2}$$

where H and N denotes the mean curvature and inward unit normal of S respectively.

A naive choice for $\Phi(\mathbf{x})$ is simply the mean of the squared errors between each pair of image values $I_i(\pi_i(\mathbf{x}))$ and $I_j(\pi_j(\mathbf{x}))$ at the locations $\pi_i(\mathbf{x})$ and $\pi_j(\mathbf{x})$ to which the surface point $\mathbf{x}$ projects (in each image for which $\mathbf{x}$ is a visible point on S). If we assume there are $n(\mathbf{x})$ visible image pairs, then we may express this choice as follows.

$$\begin{aligned} \Phi(\mathbf{x}) &= \frac{1}{n(\mathbf{x})} \sum_{i \neq j} \Phi_{ij}(\mathbf{x}) & (3) \\ \Phi_{ij}(\mathbf{x}) &= \Big(I_i\Big(\pi_i(\mathbf{x})\Big) - I_j\Big(\pi_j(\mathbf{x})\Big)\Big)^2. & (4) \end{aligned}$$

This choice penalizes individual point mismatches in the image data and is therefore extremely sensitive to noise and local texture. Such a measure is severely affected by specularities as well. To lessen these sensitivities, Faugeras and Keriven propose a local matching criterion that does not compare individual points, but small neighborhoods around these points instead. To do this, they utilize a normalized cross-correlation measure to compute the Φ_{ij} that are summed together as in (3):

$$\Phi_{ij}(\mathbf{x}) = 1 - \frac{\langle I_i, I_j \rangle}{\sqrt{\langle I_i, I_i \rangle \cdot \langle I_j, I_j \rangle}}(\mathbf{x}) \tag{5}$$

where

$$\langle I_i, I_i \rangle = \frac{1}{|\Delta|} \int_\Delta \Big(I_i(m_i{+}m) - \bar{I}_i(m_i) \Big)^2 dm,$$

$$\langle I_j, I_j \rangle = \frac{1}{|K_{ij}(\Delta)|} \int_{K_{ij}(\Delta)} \Big(I_j(m_j{+}m) - \bar{I}_j(m_j) \Big)^2 dm,$$

$$\langle I_i, I_j \rangle = \\ \frac{1}{|\Delta|} \int_\Delta \Big(I_i(m_i{+}m) - \bar{I}_i(m_i) \Big)\Big(I_j(K_{ij}(m_i{+}m)) - \bar{I}_j(m_j) \Big) dm.$$

The above integrals utilize the values of the image I_i within a neighborhood Δ (with area denoted by $|\Delta|$) around the point $m_i = \pi_i(\mathbf{x})$ as well as values of the image I_j within a neighborhood $K_{ij}(\Delta)$ (with area denoted by $|K_{ij}(\Delta)|$) around the point $m_j = \pi_j(\mathbf{x})$, where K_{ij} is a projective transformation (induced by the tangent plane of S at the point $\mathbf{x}$) which maps the point m_i to m_j. Note that we have been a bit imprecise in our notation[4] since K_{ij} actually depends both upon $\mathbf{x}$ as well as the unit normal $N(\mathbf{x})$ which, together, define the tangent plane of S at $\mathbf{x}$. Technically, this means that Φ is a function of both $\mathbf{x}$ and N, which introduces more complicated second-order terms into the gradient flow (2); however, equation (2) yields very similar results while avoiding these more expensive computations. The quantities $\bar{I}_i(m_i)$ and $\bar{I}_j(m_j)$ denote the mean values of I_i and I_j within the neighborhoods Δ and $K_{ij}(\Delta)$ around m_i and m_j respectively.

$$\begin{aligned} \bar{I}_i(m_i) &= \frac{1}{|\Delta|} \int_\Delta I_i(m_i{+}m)\, dm \\ \bar{I}_j(m_j) &= \frac{1}{|K_{ij}(\Delta)|} \int_{K_{ij}(\Delta)} I_j(m_j{+}m)\, dm. \end{aligned}$$

[4] This notation was chosen for a simple, easy-to-follow presentation.

2.3 Limitations of current approaches

The framework we just described poses multiframe stereo reconstruction as a variational problem where a matching score is minimized with respect to a surface. The matching score describes the similarity in the appearance of corresponding regions on different images of the same scene.

In the presence of specular reflections, however, the local appearance of image patches can change dramatically depending on whether one of them contains a specularity. Therefore, any of the methods based on a pure similarity measure will fail under these circumstances. Addressing the problem in its full generality would require modeling explicitly and estimating not only shape, but also reflectance properties and light distribution. Unfortunately, this is not possible since there are non-trivial combinations of different scenes that result in the same set of data. In other words, stereo and inverse global illumination under general reflectance models cannot be solved (in a sense there are more "unknowns" than "data"). One approach often adopted in the literature is to estimate either geometry or photometry with other means, or to make assumptions on the nature (e.g. shape or reflectance) of the scene.

We plan to take a more opportunistic route and exploit the very nature of specular reflections to avoid having to compute a full geometric-photometric model of the scene. We make the assumption that specular reflections (or "highlights") occupy a small portion of the scene and we use the fact that – by definition – highlights move with the viewpoint. Therefore, a highlight will never be present in the same point on the scene in more than one image (of course different highlights may be present in different images). In the next section we explain how to exploit this property to arrive at a simple and efficient algorithm.

3 Handling specular reflections

The normalized cross correlation measure yields a significant improvement over (4) in robustness to noise and to local minima induced by fine texture. In addition, the normalization helps to make the measure invariant to moderate variations in illumination from different viewpoints. However, it is still sensitive to specularities, since specular reflections tend to "saturate" the irradiance values in the image and are therefore not modeled very well by illumination shifts.

A straight-forward, but cumbersome, approach for dealing with specularities is simply to excise them from each scene image. This requires detailed and accurate detection of the specular regions within each image, which means that the performance of the following stereo reconstruction algorithm depends greatly upon the accuracy of the "specularity detector."

We now propose, instead, a method to deal with specularities more indirectly through a model that requires neither the explicit detection nor the removal of specular regions. Our approach involves replacing the first part (3) of the original naive model (3)-(4) in the same spirit that Faugeras and Keriven replaced the second part (4) with (5) to improve the robustness to noise. Simply stated, we propose using the median, rather than the mean, of the pair-wise matching scores $\Phi_{ij}(\mathbf{x})$ in order to obtain the overall value of $\Phi(\mathbf{x})$.

$$\Phi(\mathbf{x}) = \text{median}\big(\Phi_{ij}(\mathbf{x})\big) \tag{6}$$

The rationale behind using the median in place of the mean can be explained as follows. Specularities essentially form impulsive outlier regions within the scene images which therefore give rise to outlier matching scores when comparing the values of $\Phi_{ij}(\mathbf{x})$ for each visible pair of images. It is well known that the median is much more robust to such outliers than the mean (compare, for example, mean-filtering versus median-filtering of images with impulsive noise). To be more specific, it is clear that if we are given multiple views of a point $\mathbf{x}$ in the scene, that a given specularity (by the non-Lambertian nature of specular reflections) will only show up in a few of the views (typically just one view). In the case that $\mathbf{x}$ belongs to the true surface in the scene, it will give rise to good local intensity matching, as measured by $\Phi_{ij}(\mathbf{x})$, within pairs of images that do not contain the specularity, but extremely poor local matching between image pairs in which one image contains the specularity and the other does not. The mean value of Φ_{ij} will, on account of the latter pairs, indicate a poor overall matching score, whereas the median value will ignore these latter pairs and will therefore indicate a good overall matching score. We are therefore more likely to recognize the true surface in this variational framework in the presence of strong specularities by choosing the median. Note that if a point $\mathbf{x}$ does *not* lie on the true scene surface, then *both* the mean and the median are likely to exhibit poor overall matching scores.

4 Experiments

In this section we report some experiments in estimating the shape of the object shown in Figure 1. As it can be seen, the original data exhibit significant specular reflection, especially in the crown and face. Performing standard stereo, without paying attention to specularities, results in gross errors (Figure 2).

The algorithm presented in this manuscript improves significantly on these results. In Figure 3 we show the final results as estimated by our algorithm. As it can be seen the crown is captured correctly and so is some of the finer structure of the face. Details such as the nose are lost due to the regularity imposed in the estimated shape, and have nothing to do with the model of specular reflection (they would be lost even if the object was Lambertian). It is possible to reduce the level of smoothness imposed, but at the price of a more irregular reconstruction. We prefer to capture a slightly smoothed-out shape, and then texture-map the object, as we show in Figure 5, since the texture map usually helps conveying finer details. Finally, in Figure 4 we show a few snapshots of the evolution of the estimated shape starting from a large cube. This shows that, even though the method is intrinsically local and therefore convergence could lead to a local minimum, the algorithm is quite robust to the initialization, so no particular care must be observed in choosing an initial shape.

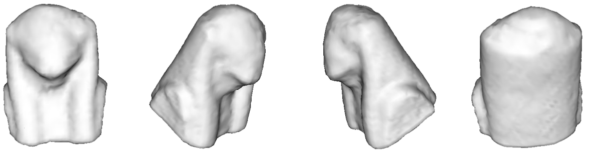

Figure 2: *Final shape estimated by a variational stereo algorithm that does not account for specular reflections. The presence of specularities on the object (Figure 1) causes outliers in the matching score that result in gross errors in the final shape. Note that the crown is missing from the reconstruction.*

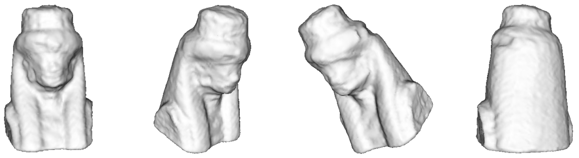

Figure 3: *Final shape estimated by our algorithm. As it can be seen, the shape of the crown is captured despite only a few views per each point being available.*

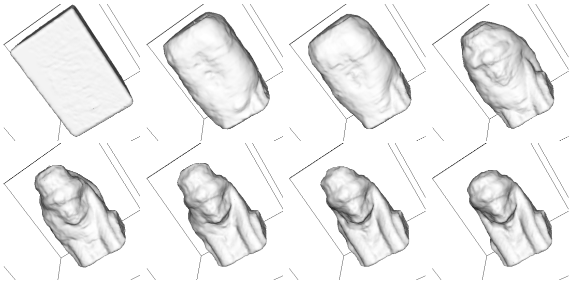

Figure 4: *Evolution of the estimate for the object in Figure 1 starting from a cube.*

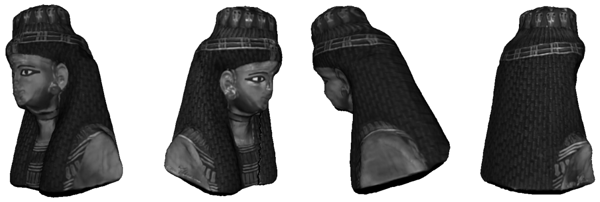

Figure 5: *Final estimate of the shape of the object in Figure 1. As it can be seen, although fine details are lost due to the smoothness term (this can be reduced at the expense of "noisier" results), the overall shape of the object is captured correctly.*

References

[1] D. N. Bhat and S. K. Nayar. Stereo in the presence of specular reflection. In *ICCV*, 1995.

[2] A. Blake. Specular stereo. In *IJCAI*, 1985.

[3] G. Brelstaff and A. Blake. Detecting specular reflections using lambertian constraints. In *ICCV*, 1988.

[4] O. Faugeras. *Three dimensional vision, a geometric viewpoint.* MIT Press, 1993.

[5] O. Faugeras and R. Keriven. Variational principles, surface evolution pdes, level set methods and the stereo problem. *INRIA Technical report*, 1996.

[6] K. Ikeuchi. Determining surface orientations of specular surfaces by using the photometric stereo method. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 3:661–669, 1981.

[7] S. K. Nayar, X. S. Fang, and T. Boult. Removal of specularities using color and polarization. In *CVPR*, pages 585–590, 1993.

[8] M. Okutomi and T. Kanade. A multiple baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(4):353–363, 1993.

[9] 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.

[10] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proc. of the AMS SIGGRAPH*, 1999.