Shape from Video: Dense Shape, Texture, Motion and Lighting from Monocular Image Streams

Azeem Lakdawalla Aaron Hertzmann

Department of Computer Science
University of Toronto
{azeem,hertzman}@dgp.toronto.edu

Abstract

This paper presents a probabilistic framework for robust recovery of dense 3D shape, motion, texture and lighting from monocular image streams. We assume that the object is smooth, Lambertian, illuminated by one distant light source, and subject to smoothly-varying rigid motion. The problem is formulated as a MAP estimation problem in which all shape, motion, noise variances and outlier probabilities are estimated simultaneously. Estimation is performed using a multi-stage initialization process followed by a large-scale quasi-Newtonian optimization technique.

1. Introduction

Traditional approaches to shape reconstruction in computer vision typically exploit either geometric cues (e.g., motion of sparse scene points) or photometric cues (e.g., change in intensity due to lighting), but not both. However, as noted recently [28, 13, 16], these cues are complementary: geometry-based methods are typically good at capturing large-scale shape for sparse informative features but not at capturing fine-scale detail or untextured regions; photometric cues are good at capturing fine-scale details, but can produce very biased estimates of global shape. Integrating these cues into a joint estimation process ought to yield methods that achieve the best of both.

This work presents a method to reconstruct dense 3D rigid geometry, motion, texture and lighting from monocular image streams captured by a freely-moving camera. Our approach is to define a generative model of image formation, and numerically optimize unknown shape and motion parameters to match the inputs. Furthermore, we employ a probabilistic formulation, allowing us to estimate model parameters (i.e., regularization factors) simultaneously with estimating shape, motion, and texture as a joint MAP estimation process. Our formulation is robust, allowing it to handle outliers such as those due to occlusions.

The imaged objects are assumed to be projected using scaled-orthographic projection, illuminated by one distant, static, light source. We require initial, sparse, point tracks of the object in order to bootstrap the system; these are currently provided by a user, but could also be provided by an automatic feature detector. We assume that the object is smooth and moves smoothly across frames.

In contrast to previous work in this area, which has employed alternating factorization steps, we directly perform MAP estimation of all unknowns by non-linear quasi-Newtonian optimization. While this procedure is very computationally demanding — requiring several days of continuous optimization for each input sequence — it is conceptually straightforward. More importantly, it recovers high quality shape reconstructions that compare favorably with existing methods.

2. Background

Perhaps the most common approach to rigid shape reconstruction is based on sparse feature correspondences. First, discriminative features are identified across an image sequence, and put into correspondence. These tracks are then fed into a structure-from-motion (SFM) algorithm that examines them and computes the shape and motion of the object. Hartley et al. [8] provide an overview of such methods.

While feature-based methods try to minimize an error metric based on tracked points, direct methods [11] minimize an error metric based *directly* on raw image data. Irani [10] exploited subspace constraints to develop a multi-point, multi-frame optical flow algorithm. Instead of tracking points individually across frames, the whole sequence, with all points, is considered at once. Torresani et al. [23] and Brand [4] use such rank constraints to overcome ambiguities in less reliable data (ie. textureless regions, occlusion, noise) and recover non-rigid geometry and motion

by using a factorization method. Torresani and Hertzmann [22] cast the same problem as an MAP estimation problem. This is done by specifying a generative model for non-rigid shape and motion based on features, where the features in turn are Lucas-Kanade [14] pixel windows. Their method is also robust to outliers. Blanz et al. [24] use *a-priori* models to fit morphable face models to images and video. This method directly minimizes the distance between image data and a projected morphable head model by summing over all pixels. Their system requires a substantial amount of training data (scanned and aligned head models) to deliver convincing reconstructions.

Shape information can also be extracted from one or more images by observing the shading variations across the imaged object's surface. Given one image of an object with constant and known albedo, illuminated by a known directional light source, shape-from-shading (SFS) [9] uses the reflectance function to determine an imaged object's normals. There is, however, an ambiguity with respect to the concave-convex nature of the shaded region [3]. Since the intensity is based on the dot product, or angle, between the normal and the light vector, there are two sets of normals that will give the same image appearance. In general, SFS algorithms do not yield convincing results [29]. Nonetheless, Zeng et al. [27] have shown that SFS can work if the user guides the process and resolves ambiguities.

Photometric stereo [26] can be thought of as an extension to shape-from-shading. By holding the viewing direction constant and varying the direction of a known light source between successive images, this technique recovers a surface's normals via the reflectance function. Belhumeur et al. [3] have shown that with three images of an object imaged under unknown lighting conditions, the surface can be reconstructed up to a Generalized Bas-Relief transformation. Basri and Jacobs [2] and Ramamoorthi and Hanrahan [19] have observed that approximately 98% of the reflected light field from a Lambertian object can be represented by the first two modes of its spherical harmonic representation. Basri and Jacobs [1] have developed a photometric stereo technique for unknown general lighting conditions based on these findings.

Recently, there has been some effort to unify these geometric and photometric approaches to reconstruction; our work falls in this category. Tracking under varying illumination is described by Jin et al. [12]. Negahdaripour [15] extends the brightness constancy constraint even further by not only adding a scaling variable but also an offset variable. Tracked patches are therefore allowed to vary linearly in intensity. Zhang et al. [28] propose to solve for dense shape and motion by a series of subspace-constrained optimizations, and Lim et al. [13] alternate factorizations and normal vector field integrations. In our work, we explore the use of direct numerical optimization of all unknowns, rather than alternative subspace optimizations, and show that the results compare favorably to existing methods. Additionally, we employ a probabilistic framework that allows us to reject outliers and to estimate noise and outlier parameters simultaneously with estimation.

3. A Generative Model for Shape From Video

We now formulate the probabilistic generative model that we use to describe the formation process for image sequences. Our model will create an image sequence of a smooth, textured, rigid, Lambertian object undergoing rotation and translation.

The geometric primitive used in our system is a height field $z = f(x, y)$. The object is imaged using scaled-orthographic projection, illuminated by one static directional light source. We assume the object undergoes smooth, rigid motion.

We define $\mathbf{p}_{i,t}$ as being the orthographically-projected, scaled, rotated, and translated surface point $\mathbf{s}_i$ at frame t:

$$\mathbf{s}_i = [x_i, y_i, f(x_i, y_i)]^T \quad (1)$$

$$\mathbf{p}_{i,t} = \rho_t \mathbf{P} \mathbf{R}_t \mathbf{s}_i + \mathbf{d}_t \quad (2)$$

where ρ_t,$\mathbf{P}$, $\mathbf{R}_t$ and $\mathbf{d}_t$ denote the scale constant, orthographic projection matrix, rotation matrix and translation vector, respectively.

We associate an RGB texture vector $[\alpha_{r_i}, \alpha_{g_i}, \alpha_{b_i}]$ to each surface point $\mathbf{s}_i$ of our height field. We incorporate shading information into our model by using the Lambertian lighting equation [6]. We also add an ambient term to approximate interreflections. The intensity value at image location $\mathbf{p}_{i,t}$, for each color channel c, is:

$$\bar{\mathbf{l}} = \frac{[l_x, l_y, 1]}{\sqrt{l_x^2 + l_y^2 + 1}} \quad (3)$$

$$I_{t,c}(\mathbf{p}_{i,t}) = \alpha_{c_i} \left(l_{a_t} + (\mathbf{R}_t \bar{\mathbf{n}}_i)^T \bar{\mathbf{l}} \right) \quad (4)$$

where $\bar{\mathbf{l}}$, $\bar{\mathbf{n}}_i$ and l_{a_t} denote the light direction vector, the normal at surface point $\mathbf{s}_i$ and the ambient light constant, respectively.

In the absence of outliers, we assume that the measured value $\tilde{I}_{t,c}(\mathbf{p}_{i,t})$ is the intensity $I_{t,c}(\mathbf{p}_{i,t})$, corrupted by additive Gaussian noise:

$$\tilde{I}_{t,c}(\mathbf{p}_{i,t}) = I_{t,c}(\mathbf{p}_{i,t}) + n \quad (5)$$

$$n \sim \mathcal{N}\left(0; \sigma_i^2\right) \quad (6)$$

where c is a color channel index (r, g, or b).

3.1. Robust Imaging Model

Each scene point has a hidden indicator variable $W_{i,t}$ associated with it. The value of $W_{i,t}$ is 0 for an outlier, or

point that does not belong to the image-generating model (such as occlusions, shadows, etc...), and 1 for an inlier. Sampling from the model entails determining $W_{i,t}$ for each point, and then sampling an image intensity, either from the outlier model (a uniform distribution) or the imaging model.

The image measurement $\tilde{I}_{t,c}(\mathbf{p}_{i,t})$ is then given by the following model:

$$p(W_{i,t}=1) = \tau \quad (7)$$

$$\tilde{I}_{t,c}(\mathbf{p}_{i,t})|I_{t,c}(\mathbf{p}_{i,t}), W_{i,t}=1 \sim \mathcal{N}\left(I_{t,c}(\mathbf{p}_{i,t}); \sigma_i^2\right) \quad (8)$$

$$\tilde{I}_{t,c}(\mathbf{p}_{i,t})|I_{t,c}(\mathbf{p}_{i,t}), W_{i,t}=0 \sim \text{Uniform}(0,1) \quad (9)$$

where c indexes over the color channels (r,g,b). Note that the outlier label $W_{i,t}$ is shared by all three color channels for each pixel.

We also assume specific prior information about the values of the unknown parameters. We assume that objects typically move smoothly from frame to frame. Objects that translate across large distances extremely quickly are very rare, and such motions should therefore be associated with small probabilities.

Smoothness can be measured by examining the second derivatives of a curve. We employ a Gaussian distribution that penalizes noisy curves using finite differences:

$$p(f(x_1),\ldots,f(x_n),\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_i (f(x_{i+1})-2f(x_i)+f(x_{i-1}))^2} \quad (10)$$

Changing the variance corresponds to changing the level of acceptable smoothness. This will be the general form of our prior probabilities. Four such terms will be added to our generative model for surface, rotation, translations and scaling.

4. The Problem Statement

Given a video sequence I of a rigid object, we would like to learn the 3D rigid shape, motion and texture of the underlying object assuming scaled orthographic projection and Lambertian illumination. In other words, we wish to obtain the parameters ζ, β, γ (described in Eqns 13, 14 and 15, respectively) that produce the highest probability $p(\zeta, \beta, \gamma|I)$. Using Bayes' rule:

$$p(\zeta,\beta,\gamma|I) = \frac{p(I|\zeta,\beta,\gamma)p(\zeta,\beta,\gamma)}{p(I)} \quad (11)$$

The denominator, or evidence, can be ignored, as it is completely independent of the unknowns ζ, β and γ. We are therefore left with the likelihood term and priors. Maximizing these together with respect to the parameters is our goal. Note that this process entails estimating shape, motion, and noise/outlier parameters simultaneously; we do not have to tune the σ^2 or τ parameters by hand.

5. Optimization

Our system uses the large scale quasi-Newton L-BFGS implementation by Zhu et al. [30], one of the best performing general optimization algorithms [17], to minimize the negative log of the probability distribution function (Eqn 17).

Our objective function is a high-dimensional manifold that contains many local minima, most of which are visually unacceptable solutions. We therefore perform three preliminary optimizations (using BFGS) to obtain variables that initialize the main optimization step. This is crucial for proper convergence.

We begin by solving for the scale, rotation, translation and shape based on selected user-tracked points across the video sequence. More specifically, we are trying to find the values of these parameters that will project these selected points onto the image plane at the same locations that the user has specified ($\mathbf{p}_{\mathbf{user}_{j,t}}$). This corresponds to a least squares optimization:

$$\begin{aligned} E_{track} = & \frac{\lambda_{track}}{2}\sum_{j,t}\|\mathbf{p}_{j,t}-\mathbf{p}_{\mathbf{user}_{j,t}}\|^2 \\ & + \frac{\lambda_{height}}{2}\sum_{j}(f(x_j,y_j))^2 \\ & + \frac{\lambda_{\text{rot}}}{2}\sum_{t}\|\text{rot}_t\|^2 \end{aligned} \quad (12)$$

where $\text{rot}_t = \begin{bmatrix} \theta_{x_{t+1}} - 2\theta_{x_t} + \theta_{x_{t-1}} \\ \theta_{y_{t+1}} - 2\theta_{y_t} + \theta_{y_{t-1}} \\ \theta_{z_{t+1}} - 2\theta_{z_t} + \theta_{z_{t-1}} \end{bmatrix}$.

The second term, which can be thought of as a prior on the shape, is there to ensure that the height field values are not too large, and will not yield a surface that is far from the XY plane. The third term is a smoothness prior on rotations. We found this necessary to reduce sporadic jumps from θ to $\theta + 2\pi$.

Once these sparse height values have been found, the rest of the surface is solved for by simply applying the surface smoothness constraint via a second optimization, but holding the points $\mathbf{p}_{\mathbf{user}_{j,t}}$ constant so that the surface passes through the user specified points.

We now have initial values for scale, rotation, translation and shape. We also require initializations for lighting, texture, τ and all σ^2 values. We now run a third optimization step to determine these values. This can be done by solving for these variables using the full objective function and holding the others variables (the ones we solved for in the previous steps) constant. This will leave us with all variables as coherent "good guesses".

Once all initialization is complete, all variables are solved for simultaneously. The algorithm proceeds in a spa-

Camera parameters:

$$\zeta = \{\mathbf{P}, \rho_{1:t}, \theta_{x_{1:t}}, \theta_{y_{1:t}}, \theta_{z_{1:t}}, t_{x_{1:t}}, t_{y_{1:t}}\} \tag{13}$$

Shape and texture parameters:

$$\beta = \{f(x_{1:t}, y_{1:t}), \alpha_{r_{1:t}}, \alpha_{g_{1:t}}, \alpha_{b_{1:t}}\} \tag{14}$$

Model and lighting parameters:

$$\gamma = \{\tau, l_{a_{1:t}}, l_x, l_y, \sigma^2_{\text{image}}, \sigma^2_{\text{shape}}, \sigma^2_{\text{scale}}, \sigma^2_{\text{rot}}, \sigma^2_{\text{trans}}\} \tag{15}$$

RGB image intensity values:

$$\kappa = \{I_{1:t,r}(\mathbf{p}_{1:i,1:t}), I_{1:t,g}(\mathbf{p}_{1:i,1:t}), I_{1:t,b}(\mathbf{p}_{1:i,1:t})\} \tag{16}$$

Complete objective function:

$$E = -\ln\left(p(I|\zeta,\beta,\gamma)p(\zeta,\beta,\gamma)\right) \tag{17}$$

$$\begin{aligned}
&= \underbrace{-\sum_{i,t} \ln\left(\tau \frac{1}{\left(2\pi\sigma^2_{\text{image}}\right)^{\frac{3}{2}}} e^{-\frac{1}{2\sigma^2_{\text{image}}}\sum_c \left(\tilde{I}_{t,c}(\rho_t \mathbf{P}\mathbf{R}_t\mathbf{s}_i + \mathbf{d}_t) - \alpha_{c_i}\left(l_{a_t} + (\mathbf{R}_t\bar{\mathbf{n}}_i)^T\bar{\mathbf{l}}\right)\right)^2} + (1-\tau)1\right)}_{\text{Image-generating term}} \\
&\quad \underbrace{+N\ln(\sigma^2_{\text{shape}}) + \frac{1}{2\sigma^2_{\text{shape}}}\sum_i \Big(\left(f(x_i+1,y_i) - 2f(x_i,y_i) + f(x_i-1,y_i)\right)^2 + \left(f(x_i,y_i+1) - 2f(x_i,y_i) + f(x_i,y_i-1)\right)^2 \Big)}_{\text{Shape prior}} \\
&\quad \underbrace{+\frac{T}{2}\ln(\sigma^2_{\text{scale}}) + \frac{1}{2\sigma^2_{\text{scale}}}\sum_t \left(\rho_{t+1} - 2\rho_t + \rho_{t-1}\right)^2}_{\text{Scale prior}} \\
&\quad \underbrace{+T\ln(\sigma^2_{\text{trans}}) + \frac{1}{2\sigma^2_{\text{trans}}}\sum_t \left(t_{x_{t+1}} - 2t_{x_t} + t_{x_{t-1}}\right)^2 + \left(t_{y_{t+1}} - 2t_{y_t} + t_{y_{t-1}}\right)^2}_{\text{Translation prior}} \\
&\quad \underbrace{+\frac{3T}{2}\ln(\sigma^2_{\text{rot}}) + \frac{1}{2\sigma^2_{\text{rot}}}\sum_t \left(\theta_{x_{t+1}} - 2\theta_{x_t} + \theta_{x_{t-1}}\right)^2 + \left(\theta_{y_{t+1}} - 2\theta_{y_t} + \theta_{y_{t-1}}\right)^2 + \left(\theta_{z_{t+1}} - 2\theta_{z_t} + \theta_{z_{t-1}}\right)^2}_{\text{Rotation prior}}
\end{aligned} \tag{18}$$

tial coarse-to-fine progression in order to avoid local minima.

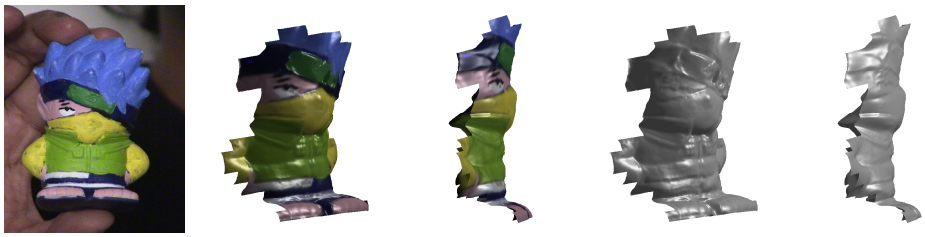
At the beginning of each step in the coarse-to-fine refinement, τ and all σ^2 values are solved for alone, as they are the only parameters that are not updated when the resolution is doubled. They must be calculated before continuing so that they match the rest of the data.

Szeliski [21] and Gortler et al. [7] have shown that convergence can be improved by representing geometry with hierarchical basis functions instead a linear finite element basis that provides only local support. Instead of representing our height field as a set of individual nodes that can move up or down, we use a wavelet basis [20] where coefficients can influence many grid points. This permits the optimization to easily make broad changes when needed.

In our system, heights fields are not the only place where such a representation can be beneficial. Since the image-matching term is dependent on both geometry and texture, we represent all three texture channels in the wavelet form as well.

6. Results

Our system has been tested on four video sequences. In each instance, a stationary camera is capturing a rigid object undergoing rotations and translations. A distant light source illuminates the object.



(a) (b) (c)

Figure 1. The Hatake Kakashi figurine. 1(a) A frame from the 234-frame sequence. 1(b) Textured reconstructions from novel viewpoints. 1(c) Untextured reconstructions from the same viewpoints.

Due to large optimization times, we were not able to operate on all images for each coarse-to-fine level. We also prematurely stop each coarse-to-fine iteration after 3000 iterations, a value found by trial-and-error. If it were possible to optimize to completion and use all frames, we expect the resulting reconstructions to have much more surface detail.

Not all experiments calculated scaling. This was "turned off" if deemed unnecessary and in situations where it was causing the surface to "shrink" significantly, thereby recovering only a portion of the intended surface. Shrinking may occur because the model can "explain", with high probability, a small patch of pixels. This drives the optimization to reduce the scaling factor so the surface projects to this small area of the image. Ambient light was not always calculated either. When it was, it was either on a per-frame basis, or as a global ambient term for the whole sequence. This was done on a trial-and-error basis.

Our system operates on blurred versions of the original images in order to reduce noise. The screenshots depicted in this section are the original un-blurred versions.

6.1. Hatake Kakashi figurine

We show results on a matte-painted figurine exhibiting very fine detail on the torso (Figure 1(a)). The entire optimization took 8 days on 1.6Ghz/1G laptop.

Figures 1(b) and 1(c) show novel views of the geometric reconstruction of the object. As can be seen, the torso contains high detail matching the input sequence, notably the lapels and the squarish buttons just below the collar.

6.2. "Lady" figurine

For comparison with recent techniques, we ran our system on the same image sequence as used in Zhang et al. [28]. Figure 2(a) shows a frame from this sequence. Figure 2(b) contains the reconstructions of those frames using the recovered shape and transformations. The entire optimization took 6 days on a 1.6GHz/1G RAM laptop.

Fine detail (spherical dimples) on the figurine's belly has clearly been reproduced. For completeness, we show a side-view comparison of the reconstructed figurine alongside a photograph taken from approximately the same angle (Figure 2(c)).

6.2.1 Comparison with other work

Figures 3(a) and 3(b) show side-view reconstructions alongside those of Zhang et al. [28]. Our results are more detailed, especially in the regions around the belly. The spherical dimples have more definition in our reconstruction.

The Phong-shaded front-view reconstruction, alongside the results from Zhang et al. [28] and Lim et al. [13], can be seen in Figure 3(c). Compared to Zhang et al. [28], our reconstruction contains much more detail, especially around the nose and mouth areas. Lim et al. [13] have reconstructed a shape that contains a few artifacts, notably on the belly where the left frontal side is indented the wrong way. This is not present in our reconstruction. Furthermore our reconstruction has better definition of the spherical dimples on the belly.

7. Uzumaki Naruto figurine

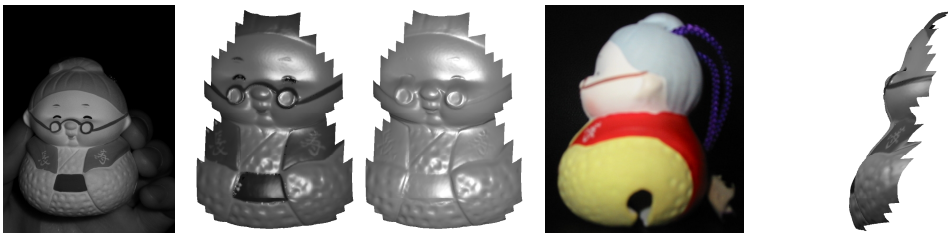
We now show results for a matte-painted figurine of Uzumaki Naruto (Figure 4). The entire optimization took 4 days on a 3.0Ghz/2G machine.

In Figures 4(b) and 4(c) we show novel views of the the recovered shape of the figurine. Once again the detail has been successfully captured by our model. The toes and fingers are evident, as are the indentations of the mouth and nose.

7.1. Occluded Uzumaki Naruto figurine

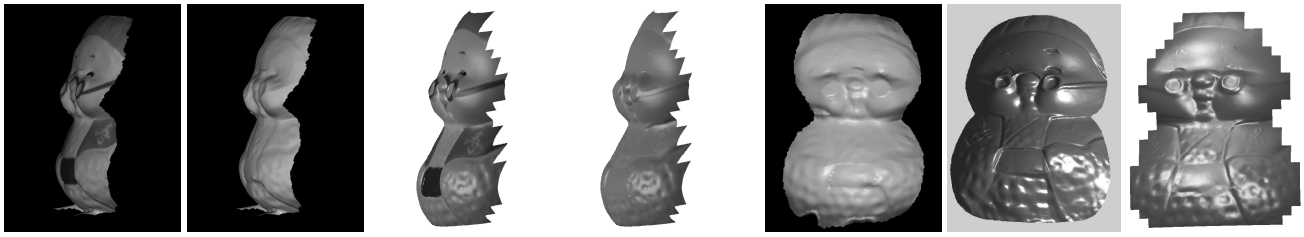
In this sequence, we occlude 30% of the frames and test our robust mixture model against a model without a mixture component for outliers (Figure 5).

The first section shows results for our robust model and the second section shows results for the model without outlier support.



(a) (b) (c)

Figure 2. The "Lady" sequence (original video footage courtesy of Zhang et al. [28]). 2(a) A frame from the 400 frame sequence. 2(b) A textured and untextured reconstruction from the same viewpoint. 2(c) A side-profile comparison of the figure and its reconstruction from the same viewpoint.



(a) (b) (c)

Figure 3. Comparisons with other work. 3(a) Textured and untextured reconstructions in Zhang et al. [28]. 3(b) Textured and untextured reconstructions from the same viewpoint using our method. 3(c) Untextured reconstructions in Zhang et al. [28] (left), Lim et al. [13] (middle) and ours (right)

7.1.1 With mixture components

The entire optimization took 4 days on a 3.0Ghz/2G machine.

Novel views of the recovered object are shown in Figures 5(b) and 5(c). Although a little noisier than the unoccluded Naruto results shown previously, we can see that the shape has nonetheless been faithfully reproduced.

It is important to note that if the optimization schedule was modified to consider more frames per coarse-to-fine level, perhaps the reconstruction would be more exact. It is also worth mentioning again that we prematurely stop each level after 3000 iterations, and it is entirely possible that more iterations may be needed in occluded circumstances.

7.1.2 Without mixture components

We now examine the recovered shape and motion using a simple Gaussian instead of our mixture model (Figure 5(d)). These experiments were conducted by forcing τ to 1.0 during the entire optimization, which took 2 days on a 3.0Ghz/1G machine.

Although the colors are generally correct, the shape is flat and contains no surface detail. Upon examining Figure 5(e), which is a reconstruction of Figure 5(a) with (left) and without (right) the mixture model, the reason for this becomes evident. When the object becomes occluded, the model "matches" the occluded frame by rotating the object out of sight in order to shade it as much as possible to simulate the hand's dark texture. Notice how the mixture model remains robust and relies on the headband to preserve orientation.

8. Discussion and Future Directions

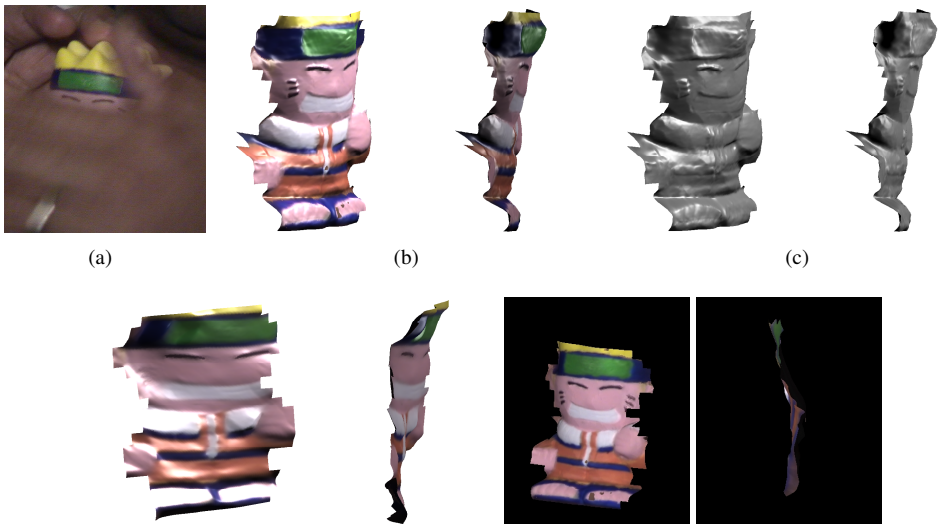
We have presented a direct, robust method for recovering dense 3D shape, texture, motion and lighting from monocular image streams. We have demonstrated that our method works, even in the presence of occlusion.

The computation time needed for recovery is quite lengthy and this poses a real bottleneck in the system. All examples took around a week of continuous calculations, and even then the optimization never reaches convergence. If we wish to reconstruct long, high resolution video sequences, it is imperative to find ways to speedup the process. The most logical decision would be to parallelize the algorithm. This is entirely possible, since the system calculates gradients by operating on one frame at a time. We could envisage distributing this task so that machines operate on different frames concurrently, and then sum all gradient calculations together before passing them to BFGS.

At present, the user manually tracks sparse features across the input sequence in order to initialize the system. If we wish to have a truly automatic system, it will be necessary to automate the feature tracking with a method that supports varying illumination [12, 15] and robustness to

(a) (b) (c)

Figure 4. The Uzumaki Naruto figurine. 4(a) A frame from the 237-frame sequence. 4(b) Textured reconstructions from novel viewpoints. 4(c) Untextured reconstructions from the same viewpoints.



(a) (b) (c)

(d) (e)

Figure 5. The occluded Uzumaki Naruto figurine. 5(a) A frame from the 203-frame image sequence. 5(b) Textured reconstructions from novel viewpoints. 5(c) Untextured reconstructions from the same viewpoints. 5(d) Textured reconstructions without using the mixture model. 5(e) Reconstruction of frame in 5(a) with and without the mixture model (left and right images respectively).

outliers [12, 22].

A logical future step would be to incorporate deformable models into the framework. This can be achieved by representing shape as a linear combination of basis shapes and placing temporal priors on shape deformations.

The constraint on Lambertian objects could also be relaxed. By considering other reflectance models [5, 25], we could reconstruct more "real world" objects since most objects are never purely Lambertian. We could simply replace the Lambertian model with another model.

There are important comparisons that need to be made with current feature-based visual modeling systems in order to determine whether shading plays a crucial role in detailed shape recovery. It is unclear, for instance, whether the system proposed by Pollefeys et al. [18] can obtain "fine-scale" detail the way photometric methods can, since most of their results are from sequences that are taken from vantage points with little illumination changes. A comparison of our and their techniques is needed on objects with fine detail in order to fully justify the importance of shading.

References

[1] R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–381, 2001. 2

[2] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. 2

[3] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999. 2

[4] M. Brand. Morphable 3d models from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 456–463, 2001. 1

[5] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics*, 1(1):7–24, 1982. 7

[6] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer graphics (2nd ed. in C): principles and practice*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1996. 2

[7] S. J. Gortler and M. F. Cohen. Hierarchical and variational geometric modeling with wavelets. In *Proceedings of the Symposium on Interactive 3D Graphics*, pages 35–42, 205, 1995. 4

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000. 1

[9] B. K. P. Horn, editor. *Shape from shading*. MIT Press, Cambridge, MA, USA, 1989. 2

[10] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *Proceedings of the International Conference on Computer Vision*, pages 626–633, 1999. 1

[11] M. Irani and P. Anandan. About direct methods. In *Proceedings of the International Workshop on Vision Algorithms*, pages 267–277, 2000. 1

[12] H. Jin, P. Favaro, and S. Soatto. Real-time feature tracking and outlier rejection with changes in illumination. In *Proceedings of the International Conference on Computer Vision*, pages 684–689, 2001. 2, 6, 7

[13] J. Lim, J. Ho, M.-H. Yang, and D. Kriegman. Passive photometric stereo from motion. In *Proceedings of the International Conference on Computer Vision*, 2005. 1, 2, 5, 6

[14] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981. 2

[15] S. Negahdaripour. Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(9):961–979, 1998. 2, 6

[16] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Papers*, pages 536–543, New York, NY, USA, 2005. ACM Press. 1

[17] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999. 3

[18] M. Pollefeys, L. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004. 7

[19] R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *Journal of the Optical Society of America A*, 18(10):2448–2458, 2001. 2

[20] E. J. Stollnitz, T. D. DeRose, and D. H. Salesin. Wavelets for computer graphics: A primer, part 1. *IEEE Computer Graphics and Applications*, 15(3):76–84, 1995. 4

[21] R. Szeliski. Fast surface interpolation using hierarchical basis functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):513–528, 1990. 4

[22] L. Torresani and A. Hertzmann. Automatic non-rigid 3d modeling from video. In *Proceedings of the European Conference on Computer Vision*, pages 299–312, 2004. 2, 7

[23] L. Torresani, D. B. Yang, E. J. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 493–500, 2001. 1

[24] T. V. Blanz, C. Basso and T. Vetter. Reanimating faces in images and video. In *Proceedings of Eurographics*, pages 641–650, 2003. 2

[25] G. J. Ward. Measuring and modeling anisotropic reflection. In *Proceedings of SIGGRAPH*, pages 265–272, 1992. 7

[26] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980. 2

[27] G. Zeng, Y. Matsushita, L. Quan, and H.-Y. Shum. Interactive shape from shading. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–350, 2005. 2

[28] L. Zhang, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. In *Proceedings of the International Conference on Computer Vision*, pages 618–625, 2003. 1, 2, 5, 6

[29] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. 2

[30] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997. 3