Shape Reconstruction from Single Relief Image

Harshit Agrawal
IIIT-Hyderabad, India
harshit.agrawal@research.iiit.ac.in

Anoop M. Namboodiri
IIIT-Hyderabad, India
anoop@iiit.ac.in

***Abstract*—Reconstructing geometric models of relief carvings are of great importance in preserving cultural heritages digitally. In case of reliefs, using laser scanners and structured lighting techniques is not always feasible or are very expensive given the uncontrolled environment. Single image shape from shading is an under-constrained problem that tries to solve for the surface normals given the intensity image. Various constraints are used to make the problem tractable. To avoid the uncontrolled lighting, we use a pair of images with and without the flash and compute an image under a known illumination. This image is used as an input to the shape reconstruction algorithms. We present techniques that try to reconstruct the shape from relief images using the prior information learned from examples. We learn the variations in geometric shape corresponding to image appearances under different lighting conditions using sparse representations. Given a new image, we estimate the most appropriate shape that will result in the given appearance under the specified lighting conditions. We integrate the prior with the normals computed from reflectance equation in a MAP framework. We test our approach on relief images and compare them with the state-of-the-art shape from shading algorithms.**

***Keywords*-Single View Reconstruction, Reliefs, Sparse Coding, Illumination Correction**

I. Introduction

Relief carvings have been a popular way of decorating buildings and depicting stories since the ancient times. They were used to enhance the ambiance of places of worship, palaces, public buildings and parks. With time many of these structures have weathered down and hence, it is of great importance to preserve these heritage symbols. Conventionally, relief surfaces are constructed by carving out stones to give an impression of a 3D shape coming out of the relief plane. The particular way of construction of reliefs provides us with useful cues that can be exploited to reconstruct the shape from a single image, which is the primary focus of this work.

Various simplifying assumptions and regularization constraints are used to solve the ill-posed problem of single view reconstruction. Still it needs strong prior knowledge about the object under consideration. An effective prior arises from the fact that surface normals at occluding contours lie in the image plane [8], [19]. However, it is very difficult to correctly detect the occluding contours in a relief, making it ineffective. We note that humans perceive shape from a single image by not only estimating the properties of the environment but also by a higher level recognition. In other words, prior knowledge from previously seen instances improves the reconstruction for us. In our approach, we encode the prior knowledge in a non-parametric way using a training database of reliefs. Shape reconstruction using exemplar database has been shown to work well in many highly similar class specific objects or shapes [2], [3] and in photometric stereo [4].

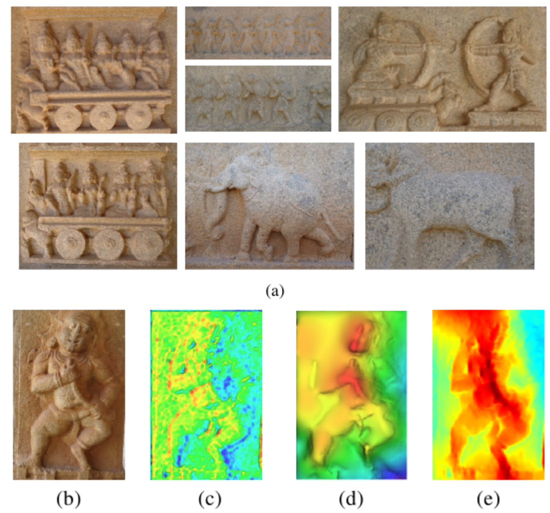


Figure 1. *Shape Reconstruction from single relief image (Depth maps are shown with psuedo-color visualization, red is near and blue is far). (a) Complete exemplar dataset consists of only 7 relief images, (b) Original Relief Image, (c) Depth Map obtained by SfS of Tsai* et al. *[20], (d) Depth Map obtained by Barron and Malik [19], (e) Depth Map obtained by our approach. The depth map obtained from (c) is noisy. We learn shape priors for reliefs to improve the shape reconstruction. Note that we recover overall geometry as well as details of face, legs and the left part of relief that is not recovered in (d). Results of [19] were poor for color images, so we used gray scale image to obtain (d).*

Techniques used for 3D reconstruction have their own limitations, especially for large scale usage. Highly accurate systems such as laser scanners are extremely expensive for use by common man, whereas multi-view stereo methods require large number of images. Other methods also make similar tradeoffs between cost, ease of use and accuracy. Our goal is to come up with an easy to use and least expensive method that improves the accuracy of relief reconstruction.

The most effective approach to recover surface normals from a single image of an object with very limited depth variation is to use the classical shape from shading (or SfS) with appropriate constraints. However, the approach assumes that we know the lighting direction or, in some cases, that it has a single frontal light source at infinity. These assumptions do not hold for images of reliefs acquired in real world as they are illuminated by a complex illumination from the environment and is rarely frontal. As described later, we overcome this challenge by using a simple modification to the imaging process using consumer cameras without the need of any additional hardware. We assume Lambertian surface reflectance model and orthographic image projection. These assumptions are valid for reliefs as the surface of reliefs are very rough and also the shape variation is very small as compared to the distance between the camera position and reliefs. The albedo, however, is not constant across the images, but is considerably uniform as the relief is made up of a single stone. We test our approach on both synthetic and real datasets. Our approach shows improvements over the shape from shading methods and is able to capture both overall shape and finer details.

Even with mostly uniform albedo of the carved reliefs, the SfS results in highly noisy depth map (see Fig. 1). We use a relief specific prior that significantly improves the results and are learned from sparse coding of sample relief images. Sparse representations of image patches are widely used for many computer vision applications like color image denoising , demosaicing and image inpainting [21]. Reliefs provide two important priors: i) the height variation across a relief is small and continuous especially in low reliefs and, ii) the overall shape of the relief is a flat plane with surface variation above the plane (see Fig. 1). Learning the relationship between the image appearance and the corresponding shape patches inherently reduces ambiguities caused by looking at an individual pixel. By using sparse representations of image patches, we are able to capture the correlation between the image appearance and local shape variations.

A vast variety of work has been done on single view shape reconstruction. Thorough and complete surveys of early work can be found in [5]. Durou *et al.* [6] surveyed recent works on numerical methods for SfS. Most of the works have popular assumptions such as Lambertian reflectance, single distant point light source, orthographic projection, and constant uniform albedo. Recent works have relaxed a few of these assumptions. Oxholm and Nishino [8] present a framework to jointly estimate the shape and reflectance of an object from single image under a known natural illumination. Similar works on shape recovery under natural illumination are Huang and Smith [9] and Jhonson and Adelson [7].

Apart from SfS approaches, researchers have examined the relationship between the shading or appearance and the shape variations in local neighborhoods [10]. Freeman *et al.* [11] presented a graphical model framework incorporating patch-based priors. In [3], database consisting of objects of highly similar class like faces, body poses etc., were used to recover the shape for a new query image of the same class. Apart from matching image appearances, they have given higher probability to patches lying in similar regions of the example images, which is possible due to the class specific database. Huang *et al.* [12] presented a generalized patch-based approach where they learn the prior probabilities for a given image patch using a database of spherical geometric primitives and their appearances. These priors are then incorporated in a variational shape from shading formulation.

II. The Proposed Approach

The core of the proposed approach involves two independent processes for estimating the shape from a relief image. The first one is based on recovery of normals using the lambertian reflection laws. Independently, we use the prior distribution of image patches from other relief images to estimate local geometric shapes. This is computed using a sparse coded representation over a relief dictionary, and serves as our relief prior for the normals. We convert surface normals to surface gradients. A MAP framework is introduced to integrate the results of the two estimates.

We also present results on a synthetic dataset of body poses, in addition to a real dataset of relief images collected from ancient heritage sites. As we see from the results, our algorithm is able to capture the overall shape and local geometric shapes, and the approach can be extended to work with objects other than reliefs.

A. Shape From Shading

Tsai and Shah [20] proposed a SfS algorithm with linear approximations that was one of the better performing algorithms in the survey by Zhang *et al.* [5]. It is a local approach where they apply discrete approximation of the gradients first, and then linearize the reflectance function in terms of the depth directly, instead of the gradients. Their approach performed well for real images, but is sensitive to noise in the intensity image. We combine their approach on the relief images along with a modified imaging process, and then improve the results by using the relief priors learned from sparse representation of query image patches.

Imaging under Known Illumination

The complex natural illumination can be simplified if we can capture two images of the same relief from approximately the same point of view, one with and the other without a flash. Flash photography is popularly used for various vision tasks such as ambient image denoising, detail transfer from flash to ambient, white balancing, red-eye correction, etc. [17]. We acquire two images of the object using a tripod to ensure the pixel alignment and to avoid image registration problem.

Let $\mathbf{A}$ be the ambient light image and $\mathbf{F}$ the image using flash. We apply gamma correction on $\mathbf{A}$ and $\mathbf{F}$ to bring both

(a) (b) (c)

Figure 2. Example of Pure Flash image computation. (a) Flash+Ambient Image **F** (b) Ambient Image **A** (c) Pure Flash Image **PF**

the images in the same linear space. Focus, aperture and ISO settings are kept same for both the images. If $\triangle t_A$ and $\triangle t_F$ are the exposure times for **A** and **F** respectively, then we compute the pure flash image **PF** as shown below

$$\mathbf{PF} = \mathbf{F} - \mathbf{A}\frac{\triangle t_F}{\triangle t_A} \tag{1}$$

Fig. 2 shows an example of computing pure flash image. Given a pair of flash and non-flash images, complex natural illumination can be simplified in the above manner to improve the accuracy and robustness of the SfS process.

In spite of the illumination correction, the SfS results are often noisy due to violations of the pure lambertian reflectance and uniform albedo assumptions of the object. We now look into the process of computing the shape prior for the image to overcome some of these problems.

B. Learning the priors for relief image

Our approach is similar in principle to the recent work from Panagopoulos *et al.* [13]. They proposed a data-driven approach that learns a dictionary of geometric primitives and their appearances. The dictionary is used to learn a small set of hypotheses about the local 3D structure for the given image to get an initial guess that is then regularized by an MRF optimization layer. In our approach, we learn the relief priors using an overcomplete dictionary with a composite signal of image appearance, surface gradients, and light source direction. To reconstruct the geometry of a given image, we sample the image densely at each pixel and for a patch around this pixel, we reconstruct a signal from the learned dictionary using a sparse linear combination of the basis signals. We use the sparse representations in learning the correlation between the image appearances and the corresponding shape variation.

Dictionary Learning: For each instance in the exemplar set, we know the gray scale image appearance I_k, surface gradients P_k and Q_k in x and y directions respectively, and the light source direction S_k. A signal $w \in \mathbb{R}^d$ in the dictionary encodes the correlation between the appearance, surface gradients and light source direction. We represent the intensity in image appearance by a square patch p (7x7 pixels) densely sampled at each pixel of I_k and surface gradients at that pixel by z_x and z_y. Each signal w is then constructed by concatenating p, S_k, z_x and z_y. Given the densely sampled signals in each instance, we learn the overcomplete dictionary as follows:

$$\{D, \alpha_i\} = \underset{D,\alpha_i}{\arg\min} \ ||w_i - D\alpha_i||_2 \quad s.t \quad ||\alpha_i||_0 < L \tag{2}$$

where D is the dictionary, w_i are the signals, α_i are the sparse representation of signals, and the constant L (L = 3) defines the required sparsity level.

For basis learning, we use the K-SVD algorithm presented in [14]. We learn the basis dictionary $D \in \mathbb{R}^{d\times n}$ where n (n = 500) is number of basis signals, such that each signal is represented by a few basis element.

Sparse Coding: Once the basis is learnt, any query signal $\mathbf{q} \in \mathbb{R}^d$ can be decomposed sparsely over the basis i.e,

$$\mathbf{q} \approx D\alpha \quad \textbf{s.t.} \quad \|\alpha\|_0 \ll L, \tag{3}$$

where α is the sparse representation of the signal and $\|.\|_0$ is l_0 pseudo-norm, which gives a measure of number of non-zero entries in a vector.

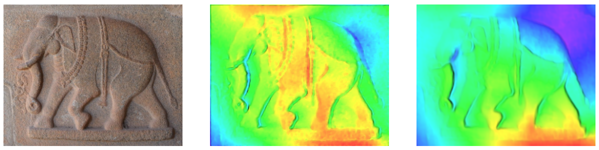
For any given image, the surface gradients are unknown. We form query signals $\mathbf{q} \in \mathbb{R}^d$ sampled densely at each pixel, with their gradient values set to zero. To represent this incomplete signal from the learned overcomplete basis, we mask the dictionary D such that the surface gradients signals are set to zero. We use the Orthogonal Matching Pursuit(OMP) technique to learn the α such that the query signal **q** is sparsely reconstructed from the basis signals. The learned α is then used to recover the corresponding surface gradient values for each pixel in the image.

C. Shape Recovery using relief priors

Given an image of a relief carving, we have now computed a shape prior and a noisy normal field from SfS. We pose the integration as a maximum-a-posteriori (MAP) estimation problem from these quantities. To achieve this, we convert normals to surface gradients and compute the most likely surface gradients $\hat{G}$ at each pixel of the image, given the observation G_s, the gradients computed from SfS. This may be written as:

$$\hat{G} = \underset{G}{\arg\max} \ p(G|G_s) = \underset{G}{\arg\max} \ p(G_s|G)p(G|G_p)$$

where G_p is the learned surface gradient priors. Note that the denominator in the Bayes formulation is not relevant for computation of $\arg\max$. The two densities, $p(G_s|G)$, and $p(G|G_p)$ models the error probabilities in the SfS and prior computations respectively. The two are estimated from ground truths of the training samples. Assuming normal distributions, the minimization has a closed form solution of the form: $\hat{G} = \alpha G_p + (1-\alpha)G_s$. α is given by $\sigma_s^2/(\sigma_s^2+\sigma_p^2)$, where σ_s^2 and σ_p^2 are the variances of the SfS and prior depth error distributions. The surface gradients thus obtained are integrated by affine transformation of gradients using diffusion tensors [16].



(a) (b) (c)

Figure 3. Comparison between two variants of signal construction discussed in Sec. III. (a) Original Image, (b) Depth map using Pixel-wise Signal Construction, (c) Depth map using Patch-wise Signal Construction. Note that, (b) captures finer details where as (c) is more smoothened shape.

III. Experiments and Results

We test our approach on relief images and synthetic datasets of Human body poses [3]. The exemplar set of relief images consists of 7 images with different lighting directions (see Fig. 1). The albedo is mostly uniform across the images with minor variations. Human body poses dataset consists of 12 images and their depth maps. For exemplar reliefs, pixel wise depth maps were computed by MVS technique. We used bundler [18] followed by dense reconstruction using PMVS [1]. The dense 3D point cloud is back projected and gaussian interpolation is used to achieve pixel wise depth correspondence.

We also test our approach with a modified signal representation. We learn the relationship between appearance and gradient patches. So, each query signal at a pixel estimates the surface gradients for a patch centered at that pixel. We refer this as patch-wise approach, and the former as pixel-wise approach. The patch-wise signal will have the following effects on prior learning. (i) Query signals are more incomplete in patch-wise, so the sparse representation may be less accurate. (ii) As each pixel will find surface gradients for a patch, the overall shape will become more smoothened and it may remove the finer geometric details. Fig. 3 shows the comparison between the two methods of signal construction.

A. Quantitative Evaluation

We use exemplar dataset to evaluate and compare our algorithms quantitatively. We learn the dictionary by leaving out the test image. Here, comparing absolute depth values is not an appropriate way of evaluating the approaches. We choose our shape evaluation metric as :

$$N - MSE(\hat{N}, N^*) = \frac{1}{n}\sum_{x,y} \arccos(\hat{N}_{x,y} \cdot N^*_{x,y})^2 \qquad (4)$$

This is the mean squared error between the angle of the normal fields $\hat{N}$ (our estimated shape) and N^* (ground-truth shape). This error metric is invariant to shifts in depth Z. Table I shows the quantitative results as average *N-MSE* for both the datasets. Our approach significantly improves upon the SfS results of Tsai *et al.* [20]. All the results were computed using a very small exemplar set with different lighting directions and we believe that our performance should improve by using a larger representative exemplar dataset or by using the modified imaging process.

	Reliefs	Human Body Poses
Tsai *et al.* [20]	0.03422	0.02817
Barron and Malik [19]	0.01868	0.01811
Our approach (Patch wise)	0.02278	0.01337
Our approach (Pixel wise)	0.02212	0.01412

Table I
Quantitative Results as Average Mean Squared Error for Reliefs and Human body poses dataset.

B. Qualitative Evaluation

In Fig. 4, we show results of our pixel-wise and patch-wise approaches on a variety of relief images captured in uncontrolled environment by consumer camera. Our pixel-wise approach performs better in 4(a), 4(c) and 4(f). We are able to recover the overall shape of the relief and also local shape variations 4(b), 4(d) and 4(e). Note that, our technique performs well in case of different lighting directions and the results can further be improved by modified imaging as discussed in Sec. II-A. Fig. 5 shows our result on body poses dataset. We correctly recovers the difference in hand positions in 5(d) and legs positions in 5(a), 5(b).

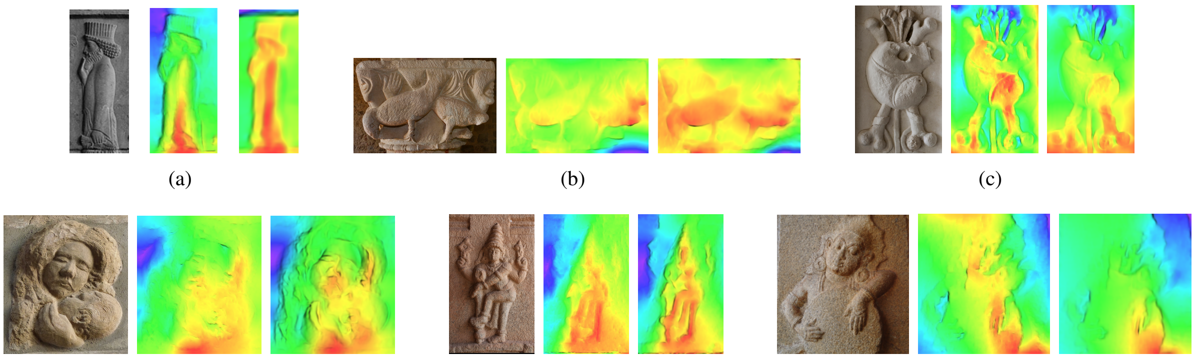
Failure Cases: Our performance is hampered in certain uncontrolled conditions. Reconstruction fails in case of cast shadows and harsh lighting conditions. We can incorporate the illumination problem with a pair of flash and non-flash images, as discussed in sec. II-A. Also our approach does not output correct shape in presence of large albedo variations. Fig. 6 shows examples for these cases.

IV. Conclusion and Future Works

We solve the shape recovery problem using a single image of relief surfaces. Reconstructing shape from relief images is a challenging task because of the uncontrolled illumination environment, so, using laser scanners or structured lighting is not always feasible. We solve the problem in two independent steps. We estimate the surface gradients using the SfS technique. The obtained gradients are noisy given the strong assumptions of SfS. We use a set of exemplar images with their corresponding shapes to learn relief specific priors. The correlation between the local image appearance and the geometric shape is learned using sparse representation technique. We remove the unnecessary complex illumination using a pair images with and without the flash. It gives us the relief image under a known illumination. After learning the relief priors, we recover the most appropriate shape by integrating the relief priors using a MAP framework. Our approach is tested on both synthetic and real datasets and result shows that our approach is able to recover both overall geometric model and local shape variations. In future, we would like to explore the modeling of non-lambertian surface reflectance to solve the shape recovery problem using sparse representations of appearance patches.

Acknowledgment

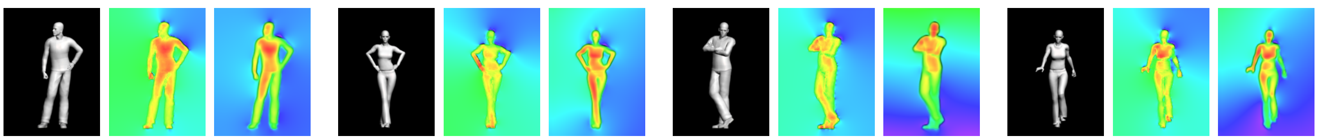
This work was partly supported by the India Digital Heritage(IDH) project of Department of Science & Technology, Govt. of India.



(a) (b) (c)

(d) (e) (f)

Figure 4. Qualitative results of our approach on relief images collected from various sources. In each instance, three images are the original image, our pixel-wise and patch-wise results, respectively. All these results are computed using the same dictionary learned on the exemplar relief images. The results shows robustness of our approach in presence of ambient illumination along with point light sources.



(a) (b) (c) (d)

Figure 5. Depth Maps obtained from our approach on Human body poses dataset [3]. Each instance is shown as original image, results of pixel-wise and patch-wise approaches as depth map respectively. The dictionary was learned using a set of 12 exemplar images. The depth variation of both the legs are correctly estimated in (a) and (b), and the depths of head in (c) and (d). Note that pixel-wise method is able to recover the depth variation of feet.

(a) (b) (c)

Figure 6. Example of failure cases. (a) Significant errors in shape reconstruction due to strong sunlight and cast shadows. (b) Incorrect shape reconstruction because of the violation of lambertian reflectance assumption. (c) Non-uniform albedo results an error in overall shape recovery.

REFERENCES

[1] Yasutaka Furukawa and Jean Ponce, *Accurate, Dense, and Robust Multiview Stereopsis*. In PAMI, 2010

[2] V. Blanz and T. Vetter, *A morphable model for the synthesis of 3d faces*. In SIGGRAPH, 1999

[3] T. Hassner and R. Basri, *Example based 3d reconstruction from single 2d images*. In CVPRW, 2006

[4] A. Hertzmann and S. M. Seitz, *Example-Based Photometric Stereo: Shape Reconstruction with General, Varying BRDFs*. In PAMI, 2005

[5] R. Zhang, P. S. Tsai, J. E. Cryer, and M. Shah, *Shape-from-shading: a survey*. In PAMI, 1999

[6] J. D. Durou, M. Falcone, and M. Sagona, *Numerical methods for shape-from-shading: A new survey with benchmarks*. In CVIU, 2008

[7] M. K. Johnson, and E. H. Adelson, *Shape estimation in natural illumination*. In CVPR, 2011

[8] G. Oxholm and K. Nishino, *Shape and reflectance from natural illumination*. In ECCV, 2012

[9] R. Huang and W. A. P. Smith. *Shape-from-shading under complex natural illumination*. In ICIP, 2011

[10] B. Potetz, and T. S. Lee. *Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes*. In JOSA, 2003

[11] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, *Learning low-level vision*. In IJCV, 2000

[12] X. Huang, J. Gao, L. Wang, and R. Yang, *Examplar-based shape from shading*. In 3DIM, 2007

[13] A. Panagopoulos, S. Hadap, and D. Samaras, *Reconstructing shape from dictionaries of shading primitives*. In ACCV, 2012

[14] M. Aharon, M. Elad, and A. Bruckstein, *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*. In IEEE Trans. Signal Process., 2006

[15] D. L. Donoho, *Compressed Sensing*. In IEEE Trans. Inf. Theory, 2006

[16] A. Agrawal, R. Raskar, and R. Chellappa, *What is the range of surface reconstructions from a gradient field?*. In ECCV, 2006

[17] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, *Digital photography with flash and no-flash image pairs*. In SIGGRAPH, 2004

[18] N. Snavely, S. M. Seitz, and R. Szeliski, *Photo tourism: exploring photo collections in 3D*. In SIGGRAPH, 2006

[19] J. T. Barron and J. Malik, *Color constancy, intrinsic images, and shape estimation*. In ECCV, 2012

[20] T. Ping-Sing, and M. Shah, *Shape from shading using linear approximation*. In Image and Vision Computing, 1994

[21] J. Mairal, M. Elad, and G. Sapiro, *Sparse representation for color image restoration*. In TIP, 2008