Reconstructing Shape from Dictionaries of Shading Primitives

Alexandros Panagopoulos[1], Sunil Hadap[2], Dimitris Samaras[1]

[1]Computer Science Dept., Stony Brook University, USA
[2]Adobe Systems Inc., USA

Abstract. Although a lot of research has been performed in the field of reconstructing 3D shape from the shading in an image, only a small portion of this work has examined the association of local shading patterns over image patches with the underlying 3D geometry. Such approaches are a promising way to tackle the ambiguities inherent in the shape-from-shading (SfS) problem, but issues such as their sensitivity to non-lambertian reflectance or photometric calibration have reduced their real-world applicability. In this paper we show how the information in local shading patterns can be utilized in a practical approach applicable to real-world images, obtaining results that improve the state of the art in the SfS problem. Our approach is based on learning a set of geometric primitives, and the distribution of local shading patterns that each such primitive may produce under different reflectance parameters. The resulting dictionary of primitives is used to produce a set of hypotheses about 3D shape; these hypotheses are combined in a Markov Random Field (MRF) model to determine the final 3D shape.

1 Introduction

Shape recovery is a classic problem in computer vision and a large body of prior work exists on the subject, including a variety of shape-from-X techniques. Shape-from-shading is the instance of the shape recovery problem where shape is inferred by the variations of shading in the image. The goal of this paper is to infer the 3D scene structure, in the form of a normal map, from a single 2D image using the information contained in shading. Although shading is a very important cue for human perception of shape and depth, shape-from-shading is a challenging and generally ill-posed problem in computer vision.

A vast amount of prior work exists in the field of shape from shading. Early work can be found in [1]. A variety of shape-from-shading algorithms are surveyed in [2], and more recently in [3], including approaches based on energy minimization and partial differential equations [4]. A variety of smoothness and curvature constraints in energy minimization is examined in [5] to improve the recovered needle maps. Energy minimization approaches suffer from deep local minima, as discussed in [6], which proposes a stochastic optimization approach to avoid them. Heavy shadows further complicate the SfS problem. In [7], shading is incorporated in the form of additional constraints to a deformable model,

in order to estimate shape under varying reflectances and extended to the case of unknown illumination. An MRF formulation of the shape from shading problem is presented in [8], including integrability constraints. While SfS is an ill-posed problem in the case of orthographic projection under a distant light source, [9] shows that assuming a more realistic perspective projection and a point light source, SfS becomes well-posed.

In our approach we are interested in extracting and utilizing information in larger image regions (*image patches*) consisting of multiple pixels. Our motivation comes from the intuition that ambiguities inherent in the problem when looking at individual pixels are reduced when examining larger neighborhoods. A data-driven approach could capture the correlations between local image appearance and geometry, allowing us to perform shape reconstruction based on a relatively small set of hypotheses about local 3D structure that have been learned by observing real data, thus making the problem easier.

Some prior work [10, 11] has examined shading and geometry in small image regions. [12] examines shading primitives capturing the shading patterns in folds and grooves of surfaces, including interreflections. A graphical model framework for incorporating patch-based priors in various computer vision problems is presented in [13]. Their results in the SfS problem are however limited to a small subset of synthetic images. Geometric primitives are also utilized in [14], to capture object-specific priors for reconstruction of known object classes, such as faces. In [15] a set of shading primitives is used to capture the folds in cloth, and the surface in between folds is interpolated through a two-level MRF model in order to reconstruct the 3D shape of cloth. Recently, [16] used learned shading primitives to deform the initially known 3D surface of a locally textured object. One of the few patch-based approaches for the general shape-from-shading problem is proposed in [17]. Their method uses a dictionary of spherical primitives and a variational approach to reconstruct the 3D shape of Lambertian objects.

In this work, we use a learned dictionary of geometric primitives to capture the relationship between the appearance and geometry of image patches. Each entry in the dictionary captures the geometry of a small rectangular region (*patch*) and a distribution of the possible image intensities associated with this geometry, as observed in a training set containing images of known geometry. We choose to describe the 3D geometry by a normal map. We assume that the scene is illuminated by a single distant point light. We do not assume a specific type of surface reflectance. In our initial approach to the problem, we assume that the object surface has uniform albedo, so that an image containing only shading variations is available. Shading variations in case of variable albedo could be extracted through other methods [18]. Furthermore, we do not model the effects of cast shadows and interreflections. However, since our method relies more on the higher-frequency components of local appearance, interreflections, which change relatively smoothly over the surface, will have limited influence on our method.

To reconstruct the shape of a new image, we first divide the image into patches. For each image patch, we search the dictionary for patches that have

similar appearance to the observed one. Patch appearance is described on a wavelet basis. We define the distance of the image patch to a dictionary patch as the Mahalanobis distance between the observed appearance and the distribution of appearances that can be produced by the dictionary patch. That distribution corresponds to different parameter choices in the Ward reflectance model [19]. Searching the dictionary for matches to an observed image patch produces a set of hypotheses about the local geometry. Despite the fact that there are infinite possible geometric explanations for the appearance of a given patch, our experiments show that certain explanations are much more probable, making our approach effective. The problem of inferring the shape of the objects in the scene becomes that of properly selecting the normal vectors given the set of local hypotheses obtained by the dictionary.

We combine the local hypotheses into the final 3D shape through a Markov Random Field (MRF) model. The MRF model contains one node per image pixel, with pairwise interactions between them, and the node labels indicate the normal vector at each corresponding pixel. The main contributions of this work are the following:

1. We propose a new metric to capture the similarity between local shading patterns and learned patches using a wavelet decomposition and the Mahalanobis distance. As a result, our method can reconstruct the shape of surfaces that significantly deviate from the lambertian model, and handle images that are not photometrically calibrated. These are both significant restrictions of previously proposed approaches.
2. We describe an algorithm that effectively combines information across multiple scales and combines the local geometric hypotheses to reconstruct the final normal map through an MRF model. Our method achieves state-of-the-art results in real images.
3. We show how a patch-based SfS approach can be used to refine and fill-in gaps in the geometry obtained with 3D sensors such as the Microsoft Kinect.

We present results on synthetic and real data. In both cases, our algorithm is able to recover both the general object shape and finer geometric details. In our experiments, dictionaries are learned on synthetic data, but we are able to use them to reliably reconstruct the shape of real photographs. Comparisons with other approaches [17, 20, 21, 9] on real data show the advantages of our approach.

In the following sections we describe how image patches can be represented and how a dictionary of patches can be learned from a set of training images and their corresponding geometry (Sec.2), and how we can reconstruct the normal map from a test image, using the trained dictionary and formulating the problem as inference on a Markov Random Field (MRF) model (Sec.3). In Sec.4 we present results on synthetic datasets and real images with our method. Sec.5 concludes the paper.

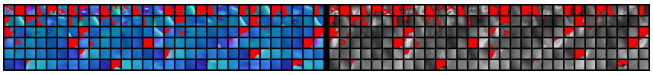

Fig. 1. The data stored in a learned dictionary. Left: the normal map of sample dictionary patches; Right: the mean appearance of each dictionary patch as reconstructed from the mean of appearance wavelet coefficients. Red indicates background pixels.

2 Patch dictionary

We first construct a dictionary of local geometric primitives (*patches*) from a set of training images with known geometry. Each patch in the dictionary is a small normal map of size $n \times n$, representing the local 3D geometry. Along with the geometry for each patch, we store the distribution of pixel intensities (local appearances) that can be produced by that geometry under different reflectance models, given a light source direction. We refer to each of the learned geometric primitives in the dictionary as a *dictionary patch*. By *patch appearance* we refer to the $n \times n$ grid of pixel intensities describing the appearance of an image patch or dictionary patch. By *patch geometry* we refer to the $n \times n$ grid of normal vectors representing the patch geometry.

2.1 Patch representation

We reduce the dimensionality of the normal map representation by applying PCA to a subset of patches from the training set and keeping the M_G first eigenvectors. Patch normal maps are therefore projected on the PCA basis and represented by the M_G resulting coefficients. We choose to represent the patch appearance using a Haar wavelet basis [22]. We use Haar wavelets of order 2, using the non-standard construction, resulting in a basis of size $M_A = 16$ for appearance patches.

The distribution of appearances that can be produced by the geometry of a dictionary patch is represented by the mean and variance of the coefficients of the patch appearance. Furthermore, each dictionary patch contains a mask that indicates which pixels belong to the foreground and which (if any) to the background. Therefore, a dictionary patch $\mathcal{D}_i$ is represented by a quadruplet $\{\mathbf{G}_i, \mathbf{M}_i, \mu_i^A, \sigma_i^A\}$, where $\mathbf{G}$ are the PCA coefficients describing the patch normal map, $\mathbf{M}_i$ is the patch foreground/background mask (an $n \times n$ grid of binary values), and μ_i^A and σ_i^A are the means and variances of the coefficients of the appearances that can be produced by the patch geometry.

An example set of patch appearances and geometries from a learned dictionary is shown in Fig.1.

2.2 Dictionary construction

Let $\mathcal{T} = \left\{\left(T_k^G, T_k^M, \mathbf{t}_k^L\right)\right\}$ be the training set, where each training instance k consists of a normal map T_k^G, a foreground/background mask T_k^M and a light source direction $\mathbf{t}_k^L$. We assume that each training instance is illuminated by a

single distant light source. In order to obtain a good dictionary $\mathcal{D}$ from training set $\mathcal{T}$, we aim to learn a set of geometric primitives that could adequately describe the objects in the training set. Our approach is to: **1)** First examine only the geometry of the training set, learning a set of dictionary patches that correspond to distinct local geometric structures in our training set. **2)** As a second step, we examine the local appearance produced by each of the learned dictionary patches under different reflectances, and store statistics to describe the distribution of these appearances.

To learn the dictionary patch geometry, we first divide the geometry T_k^G of each training instance k into a set $\mathcal{P}$ of overlapping patches P_i of size $n \times n$. We then project the normal map P_k^G of each patch P_i onto the PCA basis, so that P_k^G is represented by a set of coefficients α_k^G. To decide if we should add this patch to the dictionary $\mathcal{D}$, we compute the distance between P_k and each dictionary patch $\mathcal{D}_i$ as:

$$\langle P_k, \mathcal{D}_i \rangle = \sum_{m=1}^{M_G} \left(\alpha_k^G(m) - \alpha_i^G(m)\right)^2 + w_M \sum_{p=0}^{n^2} \left[P_k^M(p), \mathcal{D}_i^M(p)\right], \qquad (1)$$

where the first term is the euclidian distance of the PCA coefficients representing the geometry and the second term the difference of the foreground/background masks, weighed by a weight w_M that determines how strictly we want the foreground/background mask to match between the two patches (a large value of $w_M = 100$ was used in our experiments).

If the distance to the closest patch already in the dictionary is above a threshold θ_D, then a new dictionary patch is added to the dictionary, with the geometry and mask of patch P_k. Therefore, after all patches in the training set have been examined, a (potentially large) dictionary $\mathcal{D}$ has been constructed, containing a variety of distinct local geometric structures.

The second step is to learn the distribution of appearances that can be produced by the geometry of each dictionary patch. In order to do that, we render the normal map of each dictionary patch $\mathcal{D}_i$ using the Ward [19] reflectance model and a set $\mathcal{R}$ of different reflectance parameters, which corresponds to surfaces of varying specularity, varying diffuse intensity and varying anisotropic specular properties. We project the image intensities produced by each reflectance parameter selection onto the wavelet basis, and we store the mean μ_i^A and variance σ_i^A for each appearance coefficient across all reflectance parameters.

Dictionary light source direction We train the dictionary of patches using a single, known light source direction. This known light source direction is used to associate each local geometric primitive in the dictionary with a range of appearances under different reflectance parameters, removing the dependence of local appearance on light direction.

When reconstructing a test image, the light source direction used to train the dictionary has to be the same as the one that corresponds to the test image. Therefore, we re-compute the distribution of appearances for each dictionary

patch as a first step every time we are provided with a new image to reconstruct and the corresponding light source direction. Generating the distribution of appearances for a dictionary of 30000 patches, such as the one used in our experiments, takes 1-3 minutes. This time is significantly less than the time needed to reconstruct the image from the dictionary, making this solution feasible.

This way, the dictionary does not have to capture the ambiguities caused by varying light source directions, which would lead to both an extremely large dictionary and a very difficult reconstruction problem.

3 Shape reconstruction

In this section we describe how we reconstruct the geometry when provided with a new image $\mathbf{I}$ and a learned dictionary $\mathcal{D}$. We first divide the input image into a set of overlapping patches. We then find the dictionary patches in $\mathcal{D}$ that are closest in appearance to the patches extracted from the test image $\mathbf{I}$. Finally, we reconstruct the 3D shape from the results of the dictionary look-up using a Markov Random Field (MRF) model.

We divide the image $\mathbf{I}$ into a set of overlapping patches. We define an image patch P_j for each image pixel j, so that P_j is centered at pixel j and has size $n \times n$. This way, we extract all possible image patches from the input image $\mathbf{I}$. For each image patch, we search the dictionary for dictionary patches of similar appearance. We retrieve the k_D dictionary patches that are closest in terms of appearance to image patch P_j (we define the metric to compare patch appearances in the next section, Sec.3.1). Because we defined image patches centered at each pixel, a given pixel i is covered by up to n^2 overlapping image patches. As a result, there are up to $k_D n^2$ dictionary matches that include pixel i, with each dictionary match defining a normal vector for pixel i. Each of these results is considered a hypothesis about the vector at pixel i.

Because of the dependency of patches on scale, we repeat this search for a set of different scales $\mathcal{S}$. We use re-scaled versions of the original image, at scales both coarser and finer. We examine every patch at the coarsest scale. At finer scales, we only examine those image patches that have image variance above a given threshold (0.001 in our experiments). Moving to finer scales, the patches get smaller relative to the image. As a result, the average image variance per patch reduces, so that only finer details are examined at finer scales (see Fig.2). The dictionary matches of size $n \times n$ at each scale are then re-scaled to the scale of the original image. As a result, the final set of dictionary matches contains patches of varying sizes, corresponding to the different image scales used for the search.

The above procedure generates up to $|\mathcal{S}|k_D n^2$ normal vector hypotheses for each image pixel i. From this large set of hypotheses, we keep only the k normal vectors that correspond to the k dictionary patches with the lowest matching cost that contain this image pixel. These candidate normal vectors will be subsequently used in the MRF optimization described in section 3.2 to obtain the final normal map.

3.1 Dictionary search

To determine how well a dictionary patch (consisting of a normal map patch and a set of appearance statistics) matches an image patch (consisting of a patch of image intensities) we use the Mahalanobis distance.

Let P_j be an image patch consisting of appearance P_j^A (a $n \times n$ patch of per-pixel intensities) and a foreground/background mask P_j^M. Projecting the foreground pixels of appearance P_j^A onto the appearance wavelet basis, we obtain a set of coefficients α_j^A that describe the image patch appearance. We compute the distance between the appearance of P_j and that of a dictionary patch $\mathcal{D}_i$ by the Mahalanobis distance:

$$D_A(\mathcal{D}_i, P_j) = \sqrt{\sum_{m=1}^{M_A} \frac{\left(\alpha_j^A(m) - \mu_i^A(m)\right)^2}{\left(\sigma_i^A(m)\right)^2}}, \tag{2}$$

where μ_i^A and σ_i^A are the mean and variance of the appearance coefficients of the appearances produced by dictionary patch $\mathcal{D}_i$ under different reflectances, as computed during training [1].

To compute the quality of the match between dictionary patch $\mathcal{D}_i$ and image patch P_j, we also compute the similarity of the foreground/background masks of the two patches:

$$D_M(\mathcal{D}_i, P_j) = \frac{1}{n^2} \sum_{x=1}^{n} \sum_{y=1}^{n} \left[\mathcal{D}_i^M(x,y) = P_j^M(x,y)\right], \tag{3}$$

where $\left[\mathcal{D}_i^M(x,y) = P_j^M(x,y)\right] = 1$ if both masks agree for pixel (x,y) and 0 otherwise.

Finally, we can take into account the similarity of dictionary patch $\mathcal{D}_i$ to a rough 3D shape prior. This term allows us to utilize the normal map estimate from the previous scale while searching for dictionary matches at the next scale, when examining multiple scales. Similarly, this term can allow the incorporation of rough geometry knowledge. Such an example is the refinement of 3D shape captured by a commercial 3D camera, such as a Kinect sensor. The geometry prior cost is defined as:

$$D_G(\mathcal{D}_i, P_j) = \sum_{m=1}^{M} \left(\alpha_i^G(m) - \alpha_j^G(m)\right)^2, \tag{4}$$

where $\alpha_i^G(m)$ is the m-th coefficient of the geometry of dictionary patch $\mathcal{D}_j$, $\alpha_j^G(m)$ is the m-th coefficient of the *coarse* geometry of the test patch j. Assuming that the geometry prior is coarse, only the first M geometry coefficients are taken into account, corresponding to the low-frequency components of the geometry prior. In our experiments, $M = 3$.

The final cost of using dictionary patch $\mathcal{D}_i$ to explain image patch P_j is then:

$$cost(\mathcal{D}_i, P_j) = D_A(\mathcal{D}_i, P_j) + w_M D_M(\mathcal{D}_i, P_j) + w_G D_G(\mathcal{D}_i, P_j), \tag{5}$$

where w_M and w_G are weight that control the relative strength of match and geometry prior matching ($(w_M, w_G) = (1000, 1)$ in our experiments).

[1] We have assumed that covariances between appearance coefficients are 0, which lead to no significant deterioration in results, but significantly faster training and testing.

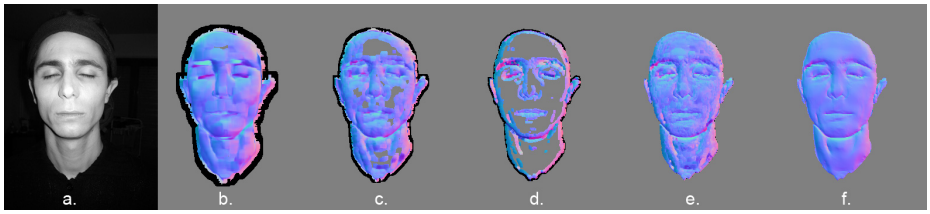

Fig. 2. Combining matches over different scales to produce an initial guess about the normal map. a) original image; b-d) the normal maps produced by averaging dictionary matches at 3 different scales; e) the combination of all scales to produce an initial guess about the normal map; f) the final result from our method.

3.2 Combination of dictionary matches

Having obtained a set of dictionary matches, we then produce an initial guess for the normal map. For each pixel i, we have recovered a potentially large set of normal vectors $\{\mathbf{n}_k^i\}$, across different scales. We compute the mean $\overline{\mathbf{n}}_i$ of all normals at pixel i. Then, we recompute the mean normals iteratively. At each iteration, we take the weighted mean of normals $\{\mathbf{n}_k^i\}$ at pixel i, where each normal is weighed by $1/||\mathbf{n}_k^i - \overline{\mathbf{n}}_i||_2$. This allows us to reduce the effect of outliers to the initial estimate [17]. The results we obtain at each scale and their combination to produce the initial guess are shown in Fig.2.

We refine this initial guess to produce the final normal map by modeling the problem as an MRF model. Through the MRF optimization, we estimate a normal map for the image that is both close to the discovered dictionary matches and that satisfies anisotropic smoothness constraints.

Our MRF model can be represented by a 4-connected 2D lattice, where each node corresponds to an image pixel. Each random variable x_i at pixel i indicates a normal vector $\mathbf{n}_i$. Therefore, the labels x_i take values from a continuous domain. The energy of the MRF model is:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{I}} \phi_i(x_i) + w_2 \sum_{i,j \in \mathcal{N}} \psi_{ij}(x_i, x_j), \tag{6}$$

where $\mathcal{I}$ is the set of image pixels, $\mathcal{N}$ is the set of neighboring pixels in the 4-connected grid, $\phi_i(x_i)$ is the singleton potential that associates the labels x_i with the geometry hypotheses recovered from the dictionary $\mathcal{D}$ and $\psi_{ij}(x_i, x_j)$ is the pairwise potential associating neighboring pixels i and j. The weight w_2 was set to 0.1 in our experiments.

The form of the *singleton potential* is:

$$\phi_i(x_i) = w_i^I \sum_{j=1}^{D_i} \arccos\left(\mathbf{n}(x_i) \cdot \mathbf{n}(\mathcal{D}_j)\right) cost(\mathcal{D}_j), \tag{7}$$

where $\mathbf{n}(x_i)$ is the normal vector at pixel i as indicated by label x_i, D_i is the number of dictionary matches that contain pixel i, $\mathbf{n}(\mathcal{D}_j)$ is the normal vector at pixel i as predicted by match $\mathcal{D}_j$, and $cost(\mathcal{D}_j)$ is the cost associated with match

$\mathcal{D}_j$. Furthermore, w_i^I is a weight that corresponds to *how reliable we expect the dictionary matches at pixel i to be*.

We express w_i^I based on two observations: dictionary matches are more reliable when there is enough local image variability (flat image regions are the least informative), and dictionary matches are not reliable when the matches in different scales differ significantly from each other. Therefore, we define w_i^I as:

$$w_i^I = \frac{\sigma_i}{1 + q(i)}, \tag{8}$$

where σ_i is the local image variance at pixel i, which is computed as the variance of the image pixel intensities in a 6×6 patch centered at pixel i. The term $q(i)$ represents how much the recovered dictionary patches differ at pixel i, and is defined as:

$$q(i) = \frac{1}{\pi} \sum_{s=0}^{|\mathcal{S}|} \sum_{j} \arccos\left(\mathbf{n}(\mathcal{D}_j^s) \cdot \overline{\mathbf{n}}_i\right), \tag{9}$$

where $\mathcal{S}$ is the set of different scales we are examining, $\mathcal{D}_j^s$ indicates the j-th recovered dictionary patch for pixel i using scale s, and $\overline{\mathbf{n}}_i$ is the normal vector at pixel i obtained by averaging the normals at pixel i from all recovered dictionary matches at all scales.

The *pairwise potentials* $\psi_{ij}(x_i, x_j)$ enforce smoothness between the normals of neighboring pixels i and j:

$$\psi_{ij}(x_i, x_j) = w_{ij} \arccos\left(\mathbf{n}(x_i) \cdot \mathbf{n}(x_j)\right), \tag{10}$$

where w_{ij} is a weight computed as a function of the image gradient between pixels i and j:

$$w_{ij} = \max\left\{0, 1 - w_\nabla \nabla I_{ij}\right\}, \tag{11}$$

and w_∇ determines how sensitive the smoothing term is to image gradients (we set $w_\nabla = 4$ in our experiments).

We infer the final shape by minimizing the MRF energy over the labels $\mathbf{x}$.We chose to use the QPBO [23, 24] and fusion-move [25] algorithms to perform inference. The QPBO algorithm is used to solve a binary MRF labeling problem between the current set of node labels $\hat{\mathbf{x}}$ and a set of proposed labels $\mathbf{x}'$. The solution is initialized to our initial guess about the normal map, produced by keeping the average normal of the finest scale available for each pixel. We perform a predefined number of iterations, and at each iteration we generate the set of proposed normals (indicated by labels $\mathbf{x}'$) by adding a small random offset to each normal vector in the current solution $\hat{\mathbf{x}}$.

4 Experimental Evaluation

We evaluated our method on both real (Fig.5) and synthetic (Fig.3) data. For evaluation on synthetic data, we used a set of 3D models rendered assuming Lambertian reflectance. The set consisted of 6 models of real objects captured with a 3D scanner [26, 27] and rendered from 142 different viewpoints and a set of 2.5D range images of 11 different objects [28], captured from 66 different

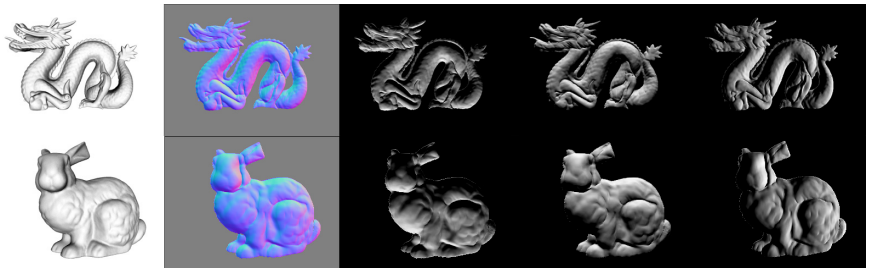

Fig. 3. Reconstruction of normal maps of synthetic images. The images are generated by rendering depth maps of objects collected by 3D scanning [26, 27]. We show the reconstructed normal maps and renderings of the reconstructed shape under different illuminations.

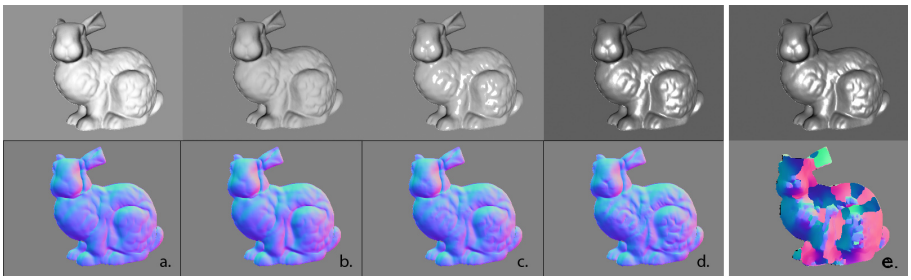

Fig. 4. Effect of non-lambertian reflectance: a-d) reconstruction using the Mahalanobis distance metric, e) reconstruction using Euclidian distance. a) Lambertian reflectance; b) Lambertian reflectance, under-exposed image; c,d,e) Specular reflectance using the Ward model. Our approach achieves results that are robust to reflectance and photometric calibration, while it is impossible to reconstruct a specular surface using just the Euclidian distance. Notice also that the surface in (d) is more specular than the most specular reflectance parameters used while training, showing the ability of our approach to generalize over reflectance parameters.

viewpoints. We used a subset of the viewpoints available, resulting in a set of 150 images. We used leave-one-out cross-validation to evaluate our algorithm: we reconstructed the shape from an image of model i using a dictionary trained on all models other than i (excluding multiple views of the same object as well). We used 4 scales (1/4, 1, 2 and 4 times the size of the original image) to recover matching patches from the dictionary. The smaller scale better captures the overall shape of the object, while finer scales can better capture detail. A total of 5000 iterations was performed during MRF inference. The running time of our algorithm was 20-40 minutes per image, depending on image size and the size of the dictionary (running time measured on an Intel Core i5 machine). Training for a dataset of 150 images takes slightly over an 1hr. We integrated the normal maps estimated by our method using the M-estimator [29], in order to produce the final 3D surfaces (Fig 6).

For our experiments, we used a dictionary of 30000 patches of size 12×12 pixels. We used a Haar wavelet basis of size 16 and the first 90 PCA eigen-

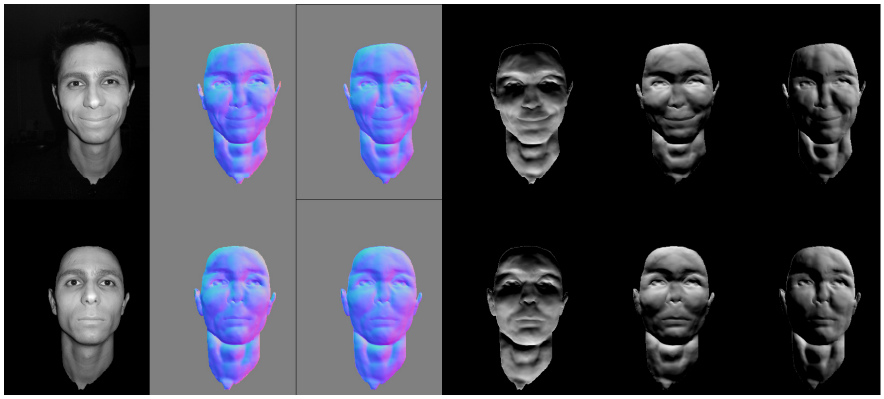

Fig. 5. Reconstruction from a real photograph. From left to right, original image (from [9]); the normal map estimated with our method; the normal map after integrating our estimate using the M-estimator [29]; 3 rendered images with the normal map we estimated and different light directions.

vectors for the patch normal maps. We observed that dictionaries of at least 10000 patches were necessary in order to get satisfactory reconstructions, while having more than 30000 patches (for the selected patch size) was usually only marginally beneficial to our results. Furthermore, it was apparent from our experiments that the patch size needs to be at least 8×8 pixels in order to properly capture local shape. We can demonstrate this through a custom dictionary containing only patches of spherical surfaces. Reconstructing an image from that dictionary is significantly more accurate with patch sizes over 8×8 pixels, which would imply that relatively large patch sizes are required to reliably capture the local curvature of surfaces, since this custom dictionary ignores finer details. Furthermore, in these experiments, using a 16×16 patch size on an image that has been rescaled to be 4 times larger than the original (without adding any detail/information) is significantly more accurate than using 4×4 patches on the original image.

In our experiments, our method significantly outperforms previous shape-from-shading approaches (Fig.7,8). It is able to reliably capture the general orientation of surfaces and is able to reconstruct much more local detail than other approaches [20, 21, 9]. This can be attributed to the fact that most shape-from-shading approaches rely on some kind of smoothness constraint, whereas in our case such constraints are replaced by the learned primitives. Smoothness needs to be enforced much more weakly during our MRF inference, allowing the solution to retain a lot of local detail. In our experiments with real data, our method also outperforms the shape-from-shading approach of [9] that applies to specific cases of the problem that can be well-posed. The ability of our method to handle surfaces that are not lambertian is one extra reason for the improved performance on real images. The use of the Mahalanobis distance further allows us to cope with images that are not photometrically calibrated (e.g. underexposed images), which can be challenging when matching the local patch appearance,

Fig. 6. Examples of 3D surfaces reconstructed from the normal maps estimated with our method, using the M-estimator [29].

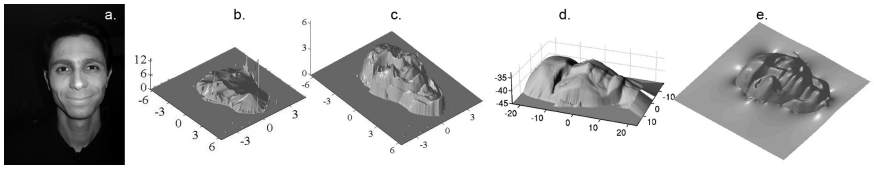

Fig. 7. Comparison of our method with other approaches: a) original image; Surface estimates by: b) [20]; c) [21]; d) [9]; e) our approach. Our approach captures both the overall shape of the object as well as the details better, resulting in a 3D face with clearly discernible features and a closer resemblance to the original face.

since in the set of reflectances used to build the distributions of appearances in the dictionary we have also included surfaces with lower uniform albedo.

One weakness of our method is that the quality of the results diminishes in the case of objects with large flat surfaces, indicating that flat patches are significantly more ambiguous than patches that contain even slight shading variations.

Refining coarse geometry We can also use our approach to refine a coarse normal map. We obtain the initial geometry using a Microsoft Kinect (a consumer device that includes a 3D scanner and a camera). The collected data are an image and a depth map. The depth values in the depth map are reliable but of low resolution. Therefore, computing the normal vectors from the depth map leads to unsatisfactory results, even when smoothing is used on the depth values, as shown in Fig.9. Furthermore, the collected depth map contains a lot of holes, especially around the occlusion borders of objects. We can use our approach to refine such results, by including the geometry information captured in Eq.5.

Fig.9 shows the results for an example scene captured using a Kinect. Our method is able to complete the holes in the collected depth map, and to obtain a convincing normal map. We show the normal maps we obtain from the Kinect depth data using various levels of smoothing on the depth values for comparison.

5 Conclusions

In this paper we presented a data-driven approach to the problem of shape-from-shading from a single image. We described how we can build a dictionary that captures the correlations between different structures in local shading and geometry. We propose a way to recover hypotheses about the local 3D geometry from the local appearance in a way that is robust to non-lambertian reflectance and photometric calibration. We recover the final 3D shape by combining these

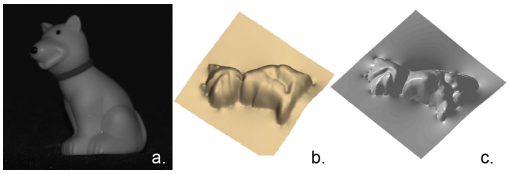

Fig. 8. Comparison of our method with [17] on a real image (from [17]): a) original image; Surface estimates: b) Result as shown in [17]; c) by our approach. Our method is able to recover more detail and a more accurate overal shape.

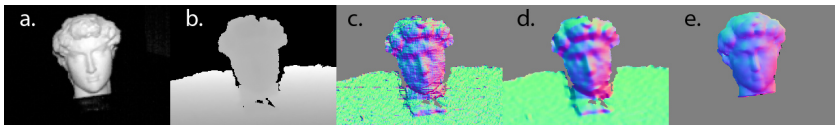

Fig. 9. Refinement of geometry captured with a Kinect: a) the image captured by the Kinect; b) the depth map captured by the Kinect; c) normals computed by the depth map; d) normals computed by the depth map after gaussian smoothing of depth values; e) normals computed by refining the smoothed normal map (d) using our method. We have correctly completed all the object edges, as well as increased the detail in the object while removing noise.

hypotheses in an MRF model. The advantages the proposed data-driven approach are that it removes a lot of typical considerations in SfS algorithms, such as boundary conditions or the choice of camera model, and enables us to explicitly deal with surfaces that deviate from the lambertian reflectance model. The results with this approach outperform previous shape-from-shading approaches, even when such approaches make significantly more assumptions than ours. The versatility of such an approach also allows us to use it in order to refine coarse geometric data captured from other sources. Future work will incorporation of priors about albedo in our dictionary representation.

Acknowledgements: This work was supported by grants NSF CNS-0627645, IIS-0916286, IIS-1111047, Adobe Systems Inc. and DIGITEO-Subsample.

References

1. Brooks, M.J.: Shape from shading. MIT Press, Cambridge, MA, USA (1989)
2. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape from shading: A survey. IEEE TPAMI **21** (1999) 690–706
3. Durou, J.D., Falcone, M., Sagona, M.: Numerical methods for shape-from-shading: A new survey with benchmarks. CVIU **109** (2008) 22–43
4. Prados, E., Faugeras, O.: A generic and provably convergent shape-from-shading method for orthographic and pinhole cameras, int. J. Computer Vision **65** (2005) 97–125
5. Worthington, P.L., Hancock, E.R.: New constraints on data-closeness and needle map consistency for shape-from-shading. IEEE TPAMI **21** (1999) 1250–1267
6. Crouzil, A., Descombes, X., Durou, J.D.: A multiresolution approach for shape from shading coupling deterministic and stochastic optimization. PAMI **25** (2003)
7. Samaras, D., Metaxas, D.: Incorporating illumination constraints in deformable models for shape from shading and light direction estimation. PAMI **25** (2003) 247–264

8. Potetz, B.: Efficient belief propagation for vision using linear constraint nodes. In: CVPR'07. IEEE Computer Society, Minneapolis, MN, USA (2007)
9. Prados, E., Faugeras, O.: Shape from shading: a well-posed problem ? In: CVPR. Volume II., IEEE (2005) 870–877
10. Potetz, B., Lee, T.S.: Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. Journal of the Optical Society of America A **20** (2003) 1292–1303
11. Potetz, B., Lee, T.S.: Scaling laws in natural scenes and the inference of 3D shape. In: NIPS 18. MIT Press, Cambridge, MA (2006) 1089–1096
12. Haddon, J., Forsyth, D.: Shading primitives: Finding folds and shallow grooves. In: ICCV. (1998) 236–241
13. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning Low-Level Vision. International Journal of Computer Vision **40** (2000) 25–47
14. T.Hassner, Basri, R.: Example based 3d reconstruction from single 2d images. In: Beyond Patches Workshop at IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society (2006) 15
15. Han, F., Zhu, S.C.: A two-level generative model for cloth representation and shape from shading. IEEE TPAMI **29** (2007) 1230–1243
16. Varol, A., Shaji, A., Salzmann, M., Fua, P.: Monocular 3d reconstruction of locally textured surfaces. PAMI (2011)
17. Huang, X., Gao, J., Wang, L., Yang, R.: Examplar-based shape from shading. In: Proceedings of the Sixth International Conference on 3-D Digital Imaging and Modeling, Washington, DC, USA, IEEE Computer Society (2007) 349–356
18. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. PAMI **27** (2005) 1459–1472
19. Ward, G.J.: Measuring and modeling anisotropic reflection. SIGGRAPH Comput. Graph. **26** (1992) 265–272
20. Falcone, M., Sagona, M.: An algorithm for the global solution of the shape-from-shading model. In: Image Analysis and Processing. Volume 1310 of LNCS. (1997) 596–603
21. Tsai, P., Shah, M.: Shape from shading using linear approximation. IVC (**12**) 487–498
22. Haar, A.: Zur theorie der orthogonalen funktionensysteme. Mathematische Annalen **69** (1910) 331–371
23. Hammer, P.L., Hansen, P., Simeone, B.: Roof duality, complementation and persistency in quadratic 0-1 optimization. Mathematical Programming **28** (1984) 121–155
24. Kolmogorov, V., Rother, C.: Minimizing nonsubmodular functions with graph cuts-a review. PAMI **29** (2007) 1274–1279
25. Lempitsky, V., Rother, C., , Blake, A.: Logcut - efficient graph cut optimization for markov random fields. In: ICCV. (2007)
26. Turk, G., Levoy, M.: Zippered polygon meshes from range images. SIGGRAPH '94, New York, NY, USA, ACM (1994) 311–318
27. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. SIGGRAPH '96, New York, NY, USA, ACM (1996) 303–312
28. Hetzel, G., Leibe, B., Levi, P., Schiele, B.: 3d object recognition from range images using local feature histograms. In: CVPR. (2001) 394–399
29. Agrawal, A., Raskar, R.: What is the range of surface reconstructions from a gradient field. In: In ECCV, Springer (2006) 578–591