Practical 3d reconstruction based on Photometric Stereo

George Vogiatzis and Carlos Hernández

Abstract Photometric Stereo is a powerful image based 3d reconstruction technique that has recently been used to obtain very high quality reconstructions. However, in its classic form, Photometric Stereo suffers from two main limitations: Firstly, one needs to obtain images of the 3d scene under multiple different illuminations. As a result the 3d scene needs to remain static during illumination changes, which prohibits the reconstruction of deforming objects. Secondly, the images obtained must be from a single viewpoint. This leads to depth-map based 2.5 reconstructions, instead of full 3d surfaces. The aim of this chapter is to show how these limitations can be alleviated, leading to the derivation of two practical 3d acquisition systems: The first one, based on the powerful Coloured Light Photometric Stereo method can be used to reconstruct moving objects such as cloth or human faces. The second, permits the complete 3d reconstruction of challenging objects such as porcelain vases. In addition to algorithmic details, the chapter pays attention to practical issues such as setup calibration, detection and correction of self and cast shadows. We provide several evaluation experiments as well as reconstruction results.

1 Introduction

Photometric stereo is a well established 3d reconstruction technique based on the powerful shading cue. A sequence of images (typically three or more) of a 3d scene are obtained from the same viewpoint and under varying illumination. From the intensity variation in each pixel one can estimate the local orientation of the surface that projects onto that pixel. By integrating all these surface orientations, a very

George Vogiatzis
Toshiba Research Europe Ltd, CB4 0GZ, e-mail: george.vogiatzis@crl.toshiba.co.uk

Carlos Hernández
Toshiba Research Europe Ltd, CB4 0GZ,e-mail: carlos.hernandez@crl.toshiba.co.uk

detailed estimate of the surface geometry can be obtained. Photometric stereo can also provide the surface reflectance properties as part of the same process.

The first reference for photometric stereo is [51]. In early papers (see [22]) the surface reflectance model was constrained to be Lambertian, an assumption that considerably simplifies calculations. Photometric stereo was subsequently relaxed to non-Lambertian reflectance models (e.g. [30, 32, 34]) but the full potential of the technique was only recently demonstrated with works such as [44, 29, 14] that obtained reconstructions of spectacular accuracy. Furthermore, in recent work [33] photometric stereo was shown to be able to significantly refine reconstruction results obtained by 3d laser range scanners. However, the method in its classic formulation suffers from some key limitations

- To obtain a reconstruction one must photograph the object in the same pose several times under changing illumination. This makes it very difficult to reconstruct a moving or deforming object during its motion.
- All photographs must be taken from a single view-point. This restricts reconstructions to 2.5D depth-maps and precludes the full reconstruction in-the-round of a closed 3d surface.

In this chapter we describe two advances in the state-of-the-art of Photometric stereo that alleviate these two limitations. Firstly we show how a coloured-light photometric stereo variant can be used to obtain independent reconstructions of the object with each photograph obtained. This makes it trivial to use the technique to reconstruct deforming objects such as moving cloth or faces. Second, we describe an elegant generalisation of photometric-stereo to multiple view-points. This method can obtain very accurate closed-surface reconstructions of objects in-the-round such as sculpture.

2 Photometric stereo with coloured light

To motivate this work, consider the problem of obtaining a dynamic 3d model of a deforming object such as cloth or a human face. This topic has received considerable attention in recent literature [38, 39, 41, 48, 49]. The complexity underlying the simplest of cloth and facial motions motivates capturing geometry and motion data from the real world.

Existing algorithms one might employ for capturing detailed 3d models of deforming objects include multiple view stereo [42], photometric stereo [20, 46], and laser based methods [28]. However, most of these techniques require that the subject stand still during the acquisition process, or move slowly [31].

This section describes a practical technique for acquiring complex motion data from real objects such as cloth or a face. The required setup consists of an ordinary video camera and three coloured light sources (see Fig. 1). The key observation is that in an environment where red, green, and blue light is emitted from different directions, a Lambertian surface will reflect each of those colours simultaneously

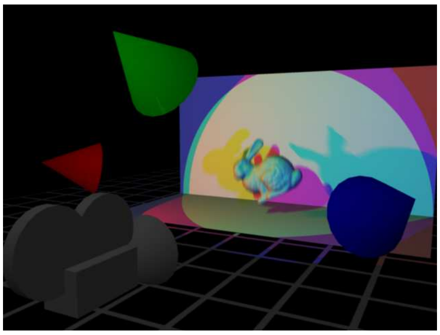


Fig. 1 Setup. A schematic representation of our multispectral setup.

without any mixing of the frequencies [37]. The quantities of red, green and blue light reflected are a linear function of the surface normal direction. A colour camera can measure these quantities, from which an estimate of the surface normal direction can be obtained. By applying this technique to a video sequence of a deforming object, one can obtain a sequence of normal maps for that object which are integrated to produce a sequence of depth-maps. In essence this technique can be seen as a variant of classic three-source photometric stereo. We will now give a brief overview of that technique and then explain how it is related to coloured photometric stereo. We will then explain in more detail some of the practical aspects of the method including calibration, how it compares numerically to ordinary photometric stereo, and how to cope with shadows.

2.1 Classic three-source photometric stereo

In classic three-source photometric stereo we are given three images of a scene, taken from the same viewpoint, and illuminated by three distant light sources. The light sources emit the same light frequency spectrum from three different non-coplanar directions. We will assume an orthographic camera (with infinite focal length) for simplicity, even though the extension to the more realistic projective case is straightforward [43]. In the case of orthographic projection one can align the world coordinate system so that the xy plane coincides with the image plane while the z axis corresponds to the viewing direction. The surface in front of the camera can then be parametrized as a height function $z(x,y)$. If ∇z is the gradient of the function wit respect to x and y, one can define the vector

$$\mathbf{n} = \frac{1}{\sqrt{1+|\nabla z|^2}} \begin{pmatrix} \nabla z \\ -1 \end{pmatrix}$$

that is locally normal to the surface at (x,y). We can also define a 2d projection operator $\mathcal{P}[\mathbf{x}] = (x_1/x_3, x_2/x_3)$ so that it follows that $\nabla z = \mathcal{P}[\mathbf{n}]$.

Now for $i = 1 \ldots 3$ let $c_i(x,y)$ denote the pixel intensity of pixel (x,y) in the i-th image. We assume that, in the i-th image, the surface point $(x,y,z(x,y))^\top$ is illuminated by a distant light source whose direction is denoted by the vector $\mathbf{l}_i$ and whose spectral distribution is $E_i(\lambda)$. We also assume that the surface point absorbs incoming light of various wavelengths according to the reflectance function $R(x,y,\lambda)$. Finally, let the response of the camera sensor at each wavelength be given by $S(\lambda)$. Then the pixel intensity $c_i(x,y)$ is given by [37]

$$c_i(x,y) = \left(\mathbf{l}_i^\top \mathbf{n}\right) \int E(\lambda) R(x,y,\lambda) S(\lambda) d\lambda. \tag{1}$$

The value of this integral is known as the surface *albedo* ρ so that (1) becomes a simple dot product

$$c_i = \mathbf{l}_i^\top \rho \mathbf{n}. \tag{2}$$

Photometric stereo methods use the linear constraints of (2) to solve for $\rho\mathbf{n}$ in a least squares sense. From this they obtain the gradient of the height function $\nabla z = \mathcal{P}[\rho\mathbf{n}]$ which is then integrated to produce the function z itself. In three-source photometric stereo, when the point is not in shadow with respect to all three lights, we measure three positive intensities c_i, each of which gives a constraint on $\rho\mathbf{n}$. If we write $L = \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 & \mathbf{l}_3 \end{bmatrix}^\top$ and $\mathbf{c} = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}^\top$ then the system has exactly one solution which is given by

$$\rho\mathbf{n} = L^{-1}\mathbf{c}. \tag{3}$$

2.2 Multi-spectral sources and sensors

This section provides the link between classic three source photometric stereo and the multi-spectral/multi-sensor case. We follow the exposition of [25]. In colour photometric stereo each of the three camera sensors (Red, Green and Blue) can be seen as a linear combination of the three images of a classic photometric stereo acquisition. To see this, consider the pixel intensity of pixel (x,y) for the i-th sensor, given by

$$c_i(x,y) = \sum_j \left(\mathbf{l}_j^\top \mathbf{n}\right) \int E_j(\lambda) R(x,y,\lambda) S_i(\lambda) d\lambda. \tag{4}$$

Note that as opposed to Eq. (1) the sensor sensitivity S_i and spectral distribution E_j are different per sensor and per light source respectively. To be able to determine a unique mapping between RGB values and normal orientation we need to assume a monochromatic surface. We therefore require that $R(x,y,\lambda) = \rho(x,y)\,\alpha(\lambda)$. Where $\rho(x,y)$ is the monochromatic albedo of the surface point and $\alpha(\lambda)$ is the characteristic chromaticity of the material. Let

$$v_{ij} = \int E_j(\lambda)\,\alpha(\lambda)\,S_i(\lambda)\,d\lambda$$

and

$$\mathbf{v}_j = \begin{pmatrix} v_{1j} & v_{2j} & v_{3j} \end{pmatrix}^\top.$$

Then the vector of the three sensor responses at a pixel is given by

$$\mathbf{c} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 & \mathbf{l}_3 \end{bmatrix}^\top \rho \mathbf{n}.$$

Essentially each vector $\mathbf{v}_j$ provides the response measured by the three sensors when a unit of light from source j is received by the camera. If matrix $\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix}$ is known, then we can compute

$$\hat{\mathbf{c}} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix}^{-1} \mathbf{c}.$$

The values of $\hat{\mathbf{c}}$ can be treated in exactly the same way as the three gray-scale images of section 2.1. The next section describes a simple process for calibrating colour photometric stereo.

2.3 Calibration

In [19] the authors propose a simple scheme for calibrating objects that can be flattened and placed on a planar board. The system detects special patterns on the board, from which it can estimate its orientation relative to the camera. By measuring the RGB response corresponding to each orientation of the material they estimate the entire matrix

$$M = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 & \mathbf{l}_3 \end{bmatrix}^\top$$

that links the normals to RGB triplets. Here we propose a two-step process. Firstly, we use a mirror sphere to estimate light directions $\mathbf{l}_1$, $\mathbf{l}_2$and $\mathbf{l}_3$. This is a standard process which has been applied in a number of photometric stereo methods. Secondly, we capture three sequences of the object moving in front of the camera. In each sequence, we switch on only one of the three lights at a time. In the absence of noise and if the monochromatic assumption was satisfied, the RGB triplets we acquired would be multiples of $\mathbf{v}_j$ when light j was switched on. We therefore do a least squares fit to the three sets of RGB to get the directions of $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. To get the relative lengths of the three vectors we can use the ratios of the lengths of the RGB vectors. The length of $\mathbf{v}_j$is set to the maximum length in RGB space, of all the triplets when light j was switched on.

2.4 Comparison with photometric stereo

To evaluate the accuracy of the per-frame depth-map estimation we reconstructed a static object (a jacket) using classic photometric stereo with three images each taken under different illumination. The same object was reconstructed using a single image, captured under simultaneous illumination by three coloured lights, using our technique. Figure 2 shows the two reconstructions side by side. The results look very similar and the average distance between the two meshes is only **1.4%** of the bounding box diagonal. It is worth noting that even though photometric stereo achieves comparable accuracy, it cannot be used on a non-static object whose shape will change while the three different images are captured. Since our method only uses one image, it is suitable for obtaining frame-by-frame reconstructions of a deforming object.

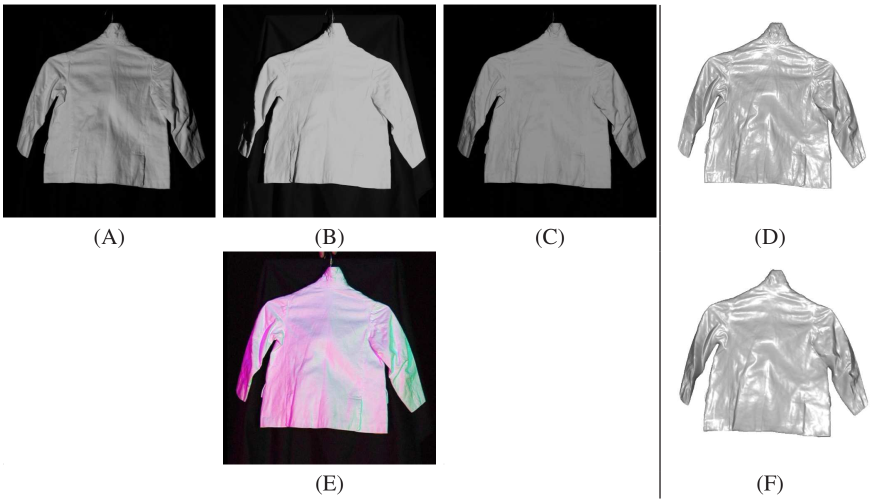


Fig. 2 Comparison with photometric stereo. (A-C) show three grayscale images captured by a digital camera, each taken under a different illumination, providing the input to a classic photometric stereo reconstruction [51] shown in (D). (E) shows a frame from a jacket sequence, where the same object is illuminated *simultaneously* by three different coloured lights. Our algorithm only uses one such frame to generate the surface mesh shown in (F). Note that both algorithms give very similar results, but only the new one (bottom row) can be applied to video since only one frame is required to obtain a reconstruction. As a quantitative comparison, the average error between both reconstructions is only **1.4%** of the bounding box diagonal.

2.5 The problem of shadows

One of the most important challenges for all photometric reconstruction methods is the frequent presence of shadows in an image. No matter how careful the arrangement of the light sources, cast or self shadows are an almost unavoidable phenomenon, especially in objects with complex geometries. This section examines in detail the phenomenon of shadows in photometric stereo with three light sources.

Shadows in photometric stereo have been the topic of a number of papers [2, 6, 11]. Most papers assume we are given four or more images under four different illuminations. This over-determines the local surface orientation and albedo (3 degrees of freedom) which implies that we can use the residual of some least squares solution, to determine whether shadowing has occurred. However when we are only given three images, as in the case of colour photometric stereo there are no *spare* constraints against which to test our hypothesis. Therefore the problem of detecting shadows becomes more difficult. Furthermore, when a pixel is in shadow in one of the three images most methods simply discard it. Here we show how one can use the remaining two image intensity measurements to estimate the surface geometry inside the shadow region. Using an argument based purely on counting degrees of freedom and equations, this is theoretically possible since we need to estimate 2 DOF per pixel (depth and albedo) and we have two independent measurements per pixel. The solution is based on enforcing (1) integrability of the gradient field, as well as (2) smoothness in the recovered shape.

Consider a three-source photometric stereo setup where one point is in shadow, say in the 3-rd image. This implies that the image measurement of c_3 cannot be used as a constraint. Since each equation (2) describes a 3d plane, the intersection of the two remaining constraints is a 3d line given by

$$(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)^\top \mathbf{n} = 0, \tag{5}$$

or equivalently

$$\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]^\top \nabla z = 1. \tag{6}$$

This equation was derived by [50] and used for stereo matching in a two-view photometric stereo setup, and subsequently used by [12] to perform uncalibrated photometric stereo and by [7] in their proof of non-existence of a general illumination invariant. Here we show how this equation can be used in a least squares framework to perform three-source photometric stereo in the presence of shadows.

2.5.1 Integrating in the shadowed regions

According to the image constraints and assuming no noise in the data, we can have one of the following three cases:

1. The surface point is in shadow in two or more images. In this case there is no constraint in ∇z from the images.

2. The surface point is not in shadow in any of the three images. In this case ∇z coincides with $\mathcal{P}\left[L^{-1}\mathbf{c}\right]$.
3. The surface point is in shadow in exactly one image, say the 3rd. In this case ∇z must lie on the line $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]^\top \nabla z = 1$. We call this line the *shadow line* of the shaded pixel.

Now in the presence of noise in the data $\mathbf{c}$, cases 2 and 3 above do not hold exactly as $\mathcal{P}\left[L^{-1}\mathbf{c}\right]$ and $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]$ are corrupted. The estimation of the unknown height function z becomes a least squares problem with two different data terms, one for pixels under shadow and another one for pixels seen in all three images.

Under noise in the image data $\mathbf{c}$, the 2d point $\mathcal{P}\left[L^{-1}\mathbf{c}\right]$ and 2d line $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]$ are not perfectly consistent with the model. For non-shadowed pixels, the difference between model and data can be measured by the *point-to-point* square difference term

$$\mathcal{E} = |\nabla z - \mathcal{P}\left[L^{-1}\mathbf{c}\right]|^2. \tag{7}$$

In the case of the shadowed pixels we have a *point-to-line* square difference term

$$\overline{\mathcal{E}}^{(3)} = (\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]^\top \nabla z - 1)^2. \tag{8}$$

Assume we are given a labelling of pixels into all the possible types of shadow. Let $\mathcal{S}$ contain all non-shadowed pixels while $\mathcal{S}_i$ contains pixels shaded in the i-th image. Our cost function becomes

$$\sum_{j\in\mathcal{S}} \mathcal{E}_j + \sum_{j\in\mathcal{S}_1} \overline{\mathcal{E}}_j^{(1)} + \sum_{j\in\mathcal{S}_2} \overline{\mathcal{E}}_j^{(2)} + \sum_{j\in\mathcal{S}_3} \overline{\mathcal{E}}_j^{(3)}$$

which is a set of quadratic terms in ∇z and thus z. Finding the minimum of this quantity is a simple unconstrained linear least squares problem that can be solved using a sparse linear solver such as UMFPACK [9].

Figure 3 shows this idea applied in practise on synthetic data. It provides evidence that in its present form the problem is ill-conditioned, especially in larger shadowed regions (see Fig. 3c). The following section sheds more light on this and describes our proposed remedy (see figures 3d and 3e).

2.5.2 Regularisation in the shadow regions

The linear least squares optimisation framework described in section 2.1 when executed in practise shows signs of ill-posedness in the presence of noise. This is demonstrated in the synthetic case of figure 3 where three images of a sphere have been generated. Three shadow regions corresponding to each of the three lights have been introduced. Even though the overall shape of the object is accurately captured, some characteristic ‘scratch’ artifacts are observed. These are caused by the point-to-line distances which do not introduce enough constraints in the cost function. The point ∇z can move significantly in a direction parallel to the corresponding shadow line only to gain a slight decrease in the overall cost. This results in violent perturbations in the resulting height function that manifest themselves as deep scratches that follow the 2d flow $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]$.

If we push the analysis even further and have one of the images completely shadowed, we then fall back to the two-source photometric stereo setup shown in Fig. 4. When only two images are available without shadow (see Fig. 4 top), after factoring out the albedo (5) we can only determine the depth gradient along specific directions for each pixel $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]$. If we look at these directions as a vector field, then depth can be computed independently along each streamline or "characteristic curve" (see Fig. 4b). In other words, there is no constraint between the depth of two characteristic curves and one pixel can only belong to a single characteristic curve. After integrating every characteristic curve independently (see Fig. 4**c**), we obtain a possible reconstruction that is different from the original true shape, but that perfectly agrees with the given constraints. In order to choose one among the possible solutions, some type of regularisation is needed (see Fig. 4d and 4e).

Regularisation can be seen as a prior on the type of solutions we expect. In order to better understand what types of prior might be relevant, let us restate the problem assumptions. We have a three source photometric stereo setup with varying albedo, *and* one of the lights is occluded, *i.e.* we locally have a two source photometric stereo setup with varying albedo. From the theory we know that in the photometric stereo setup, the albedo and the geometry are coupled, and if there is not enough data available, both are indistinguishable. This coupling exactly indicates what two types of priors one might use: either a shape smoothness prior favouring smooth shapes or an albedo smoothness prior favouring smooth albedo. The exact type of prior used should depend on the type of data captured. A good regularising criterion must satisfy two main requirements:

- The scheme must be consistent with the linear least squares framework. No nonlinear constraints can be enforced.
- It must suppress noise while preserving as much of the data as possible.

In the following we describe two different regularisation schemes that favour smooth shapes while preserving the data as much as possible. Their main difference is that one favours shapes with a *smooth shading* under the occluded light, while the second one favours *smooth shapes*. The second scheme can be used in a two-source photometric stereo setup as it is independent of the occluded third light (see Fig. 4).

Shading regularisation

In this approach we want to impose regularisation on the collected shading intensities, thereby ”inpainting” [4] the shadowed regions in order to recover the intensities we would collect had the light not been occluded and the albedo been constant. From equation (3) we can parametrize the shadow line as a function of the missing shading μ

$$\nabla z = \mathcal{P}\left[L^{-1}\begin{pmatrix}c_1\\c_2\\0\end{pmatrix} + \mu\rho L^{-1}\begin{pmatrix}0\\0\\1\end{pmatrix}\right] \tag{9}$$

This parameter represents the value $\mathbf{l}_3^\top \mathbf{n}$ would have, had the point not been in shadow in the 3-rd image. In order to simplify the notation of (9) we define matrix $M = L^{-1}$ where vector $\mathbf{m}_i$ is the i^{th} column of matrix M, giving

$$\nabla z = \mathcal{P}[c_1\mathbf{m}_1 + c_2\mathbf{m}_2 + \mu\rho\mathbf{m}_3] \tag{10}$$

We observe that, because c_1 and c_2 already encode the albedo ρ in, equation (10) is in fact independent of ρ due to the projection operator. We also note that ∇z is not a linear function of μ meaning that we cannot directly regularise the missing shading μ in a linear least squares framework. However, we can perform a change of variables and introduce a new variable w per shaded pixel

$$w(\mu) = \frac{\mathbf{e}_3^\top(c_1\mathbf{m}_1 + c_2\mathbf{m}_2)}{\mathbf{e}_3^\top(c_1\mathbf{m}_1 + c_2\mathbf{m}_2) + \mu\rho\mathbf{e}_3^\top\mathbf{m}_3}, \tag{11}$$

with $\mathbf{e}_3 = (0,0,1)^\top$. The new variable w still specifies a location along the shadow line of that pixel so equation (10) simply becomes

$$\nabla z = w\mathcal{P}[c_1\mathbf{m}_1 + c_2\mathbf{m}_2] + (1-w)\mathcal{P}[\mathbf{m}_3] \tag{12}$$

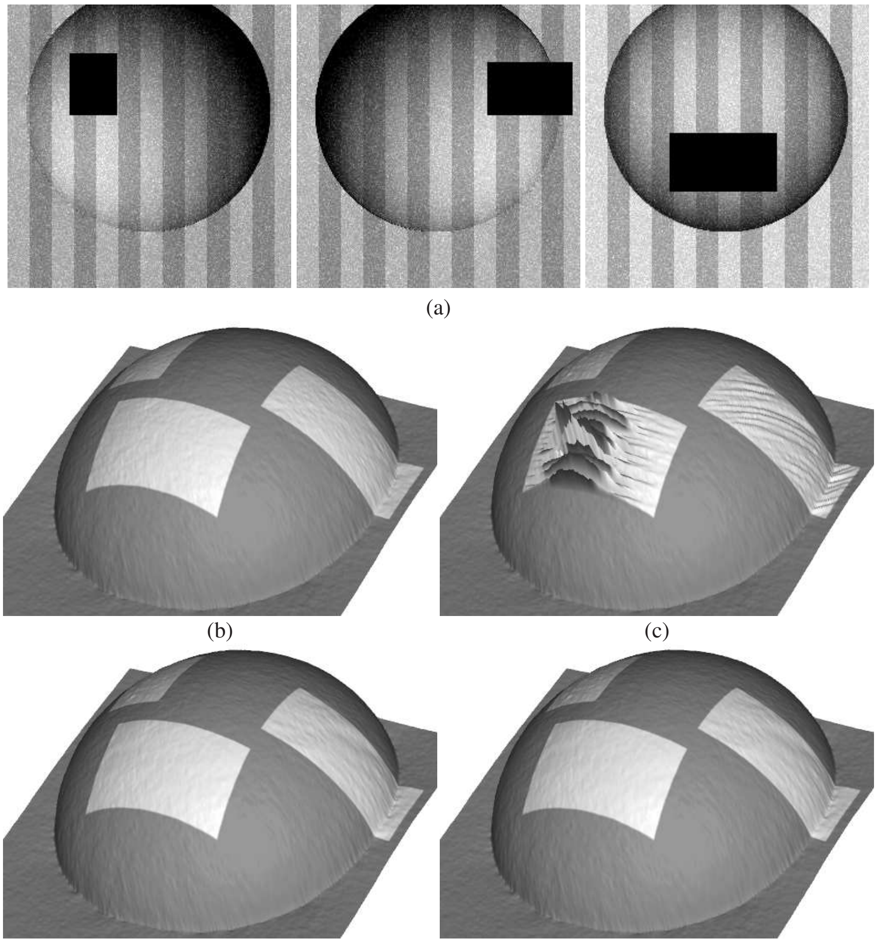
The term is now quadratic with respect to ∇z and w, allowing us to regularise the solution in a meaningful way by using first order $|\nabla w|$ and second order $|\nabla^2 w|$ regularisation terms on w. The point-to-line distance of (8) can now be replaced with the following point-to-point distance

$$\begin{aligned}\overline{\mathcal{E}}^{(3)} = {} & |\nabla z - w\mathcal{P}[c_1\mathbf{m}_1 + c_2\mathbf{m}_2] - (1-w)\mathcal{P}[\mathbf{m}_3]|^2 + \\ & \alpha|\nabla w|^2 + \beta|\nabla^2 w|^2,\end{aligned} \tag{13}$$

where α and β are regularisation weights. As w is a proxy for μ, this corresponds to introducing smoothness in the product $\mathbf{l}_3^\top \mathbf{n}$. We can therefore eliminate the scratch artifacts while letting $\mathbf{n}$ have variability in the directions perpendicular to $\mathbf{l}_3$.

Shape regularisation

The most common way of regularising shape is using first-order and second-order regularisation terms. In the context of a height field, this is achieved by minimising the norm of the gradient of the height field $|\nabla z|$ or minimising the Laplacian of the height field $|\nabla^2 z|$. The latter is known to have good noise reduction properties and to produce smooth well behaved surfaces with low curvature. However, both the gradient and the Laplacian are isotropic so they tend to indiscriminately smooth along all possible directions. See [1] for a good discussion of anisotropic alternatives to Laplacian filtering in the context of gradient field integration. In the context of our



(a)

(b) (c)

(d) (e)

Fig. 3 Regularization schemes. This is an experiment on a synthetic sphere designed to validate the proposed regularisation constraints. **(a)** shows the input images where the black rectangles correspond to occluded regions.This object is illuminated from three directions and the three white regions are occluded in the corresponding images. Middle row shows the photometric stereo solution without shadows **(b)** and the effect of optimising the surface with no regularisation at all,*i.e.* just using integrability **(c)**. Note the characteristic ‘scratch’ artifacts. **(d)** shows the resulting surface after adding a shading regularisation term with optimal values $\alpha = 6.1, \beta = 0$. **(e)** shows the resulting surface after adding a shape regularisation term with optimal values $\alpha = 0.08, \beta = 0.46$. See Section 2.5.2 for a description of the algorithms. The artifacts have been suppressed while the data has been preserved unsmoothed. Note how both regularisation schemes give almost identical results.

problem, there is an efficient way of achieving anisotropic versions of both the first-order and the second-order regularisation terms. From equation (6), we observe that the shape is totally unconstrained along perpendicular directions to $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]$. The directions $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]$ define characteristic curves, visually showing the constraint induced by the two lights (see Fig. 4b). Therefore a good way of regularising the shape is along perpendicular directions $\mathbf{u}$ to the characteristic curves, *i.e.* $\mathbf{u}^\top\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)] = 0$. The point-to-line distance term (8) is therefore extended with anisotropic first and second order regularisation terms

$$\begin{aligned}\overline{\mathcal{E}}^{(3)} = &(\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]^\top \nabla z - 1)^2 \\ &+\alpha|\mathbf{u}^\top\nabla z|^2 + \beta|\mathbf{u}^\top H(z)\mathbf{u}|^2,\end{aligned} \tag{14}$$

α and β being the regularisation weights and $H(z)$ the Hessian matrix.

Throughout all of the previous discussion we have assumed knowledge of labelling of pixels according to shadows. The next section discusses how we propose to segment shadow regions in the image.

2.5.3 Segmenting shadowed regions

It is known [2] that in photometric stereo with four or more images one can detect shadows by computing the scaled normal that satisfies the constraints in a least squares sense. If the residual of this least squares calculation is high, this implies that the pixel is either in a shadow or in a highlight. With three images however this becomes impossible as the three constraints can always be satisfied exactly, leaving a residual of zero. Recently, [6] proposed a graph-cut based scheme for labelling shadows in photometric stereo with four or more images. Based on the constraint residual, they compute a cost for assigning a particular shadow label to each pixel. This cost is then regularised in an MRF framework where neighbouring pixels are encouraged to have *similar* shadow labels. We would like to use a similar framework but we must supply a different cost for assigning a shadow label. The basic characteristic of a shadow region is that pixel intensities inside it are dark. However this can also occur because of dark surface albedo. To remove the albedo factor we propose to divide pixel intensities with the magnitude of the intensity vector $\mathbf{c}$. Our cost for deciding that a pixel is occluded in the i-th image is $c_i/\,\|\mathbf{c}\|$. This still leaves the possibility that we mistakenly classify a pixel whose normal is nearly perpendicular to the i-th illumination direction $\mathbf{l}_i$. However in that case the pixel is close to being in a self shadow so the risk from misclassifying it is small. The cost for assigning a pixel to the non-shadowed set is given by

$$\frac{1}{\sqrt{3}} - \min_i \frac{c_i}{\|\mathbf{c}\|}.$$

We regularise these costs in an MRF framework under a Potts model pairwise cost [15]. This assigns a fixed penalty for two neighboring pixels being given different shadow labels. The MRF is optimised using the Tree Reweighted message passing

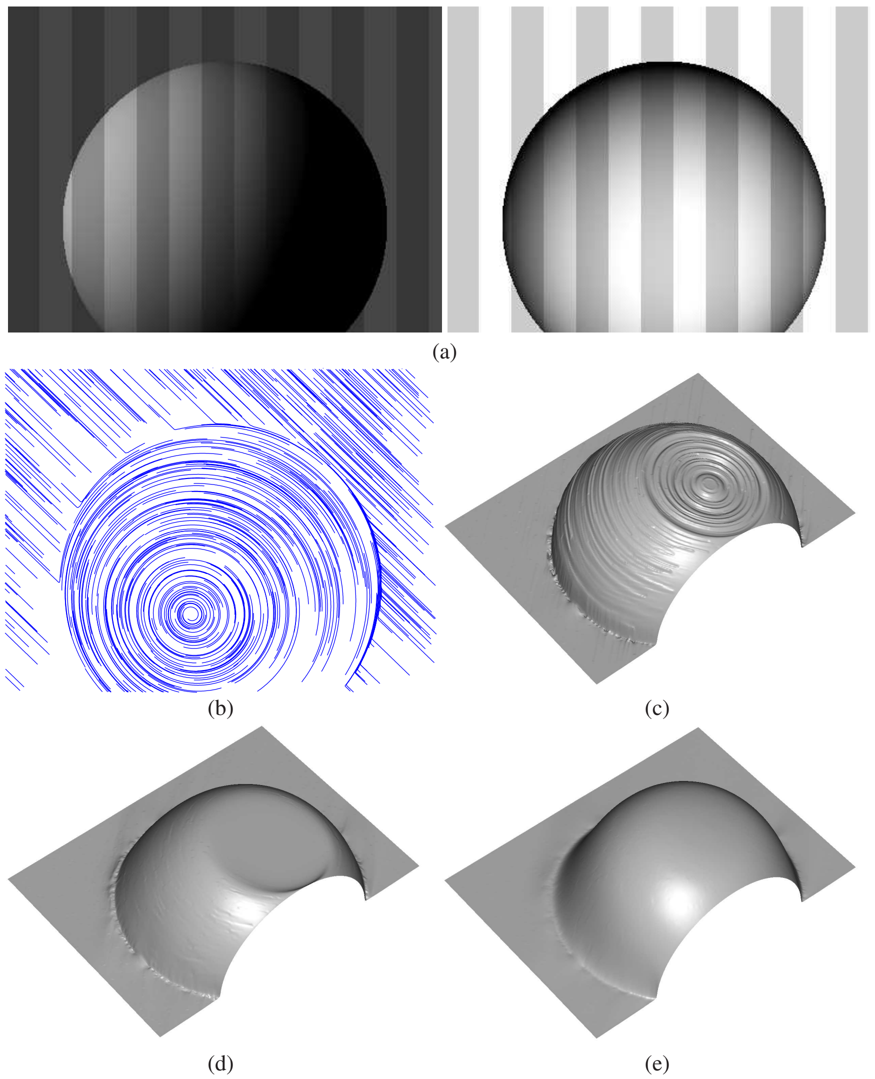


Fig. 4 Two-source varying albedo photometric stereo setup. In this experiment we show a two-source photometric stereo with varying albedo. **(a)** shows the two input images. **(b)** shows the characteristic curves obtained by plotting seeds following the 2d flow $\mathcal{P}[(c_2\mathbf{l}_1 - c_1\mathbf{l}_2)]$. **(c)** shows one possible reconstruction of the characteristic curves. Note how each characteristic curve is reconstructed independently as there is no constraint "across" the curves. Bottom row shows how a successful reconstruction can be achieved when using the proposed shape regularisation scheme with first order regularisation $\alpha = 0.1, \beta = 0$ **(d)** and second order regularisation $\alpha = 0, \beta = 0.5$ **(e)**.

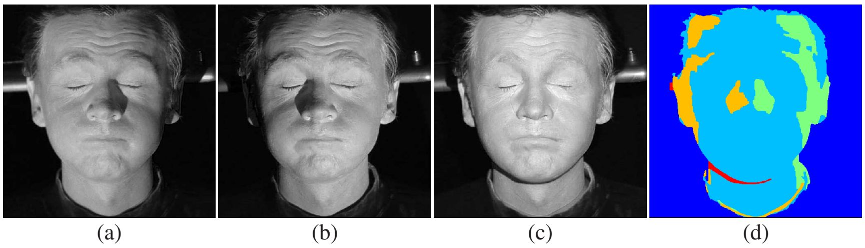


Fig. 5 Shadow segmentation. This experiment shows the result of our shadow region segmentation. From left to right, the three input images **(a), (b), (c)** and the mask with the resulting shadow labels **(d)**.

algorithm [24]. Figure 5 shows an example of applying our shadow region segmentation to a real image.

2.6 Facial Capture Experiments

We have performed a first experiment with video data of a white-painted face illuminated by three coloured lights in a similar way as in [19]. The setup is calibrated as described in section 2.3. Figure 5 shows the three input images obtained from a single colour frame. The automatic shadow segmentation results in Fig. 5d demonstrate the accuracy of the shadow detection algorithm in Section 2.5.3. Figure 6 shows three different frames of the video sequence without taking the shadows into account (left) and after detecting and adding the shading constraints (middle) and the shape constraints (right). We can appreciate how the nose reconstruction is dramatically improved when correctly processing the shadows (see arrows), even though only two lights are visible in the shadowed regions. We also note that the shape regularisation scheme fails in some boundary regions (see circles in right column) leading to an incorrect reconstruction of the side of the face. This is caused by the Laplacian regularisation term. The term suffers from an ambiguity of two possible solutions, concave or convex, both solutions having similar energy and the data term being unable to disambiguate them.

Figure 7 shows a more detailed analysis of the bottom face in figure 6. The solution of the shape regularisation scheme agrees with the constraints (Fig. 7 left) even though it picks the incorrect “concave” solution instead of the convex solution. This is confirmed by looking at the shade rendering of the face under the occluded light (see Fig. 7 middle and right). The shading regularisation scheme shows a smooth surface (Fig. 7 middle) while the shape regularisation scheme (Fig. 7 right) shows a clear artifact. This is expected since the shading regularisation does exactly that, it finds the surface that minimises the variation of the shading when rendering the shape with the occluded light. The extra knowledge of the missing light is exactly

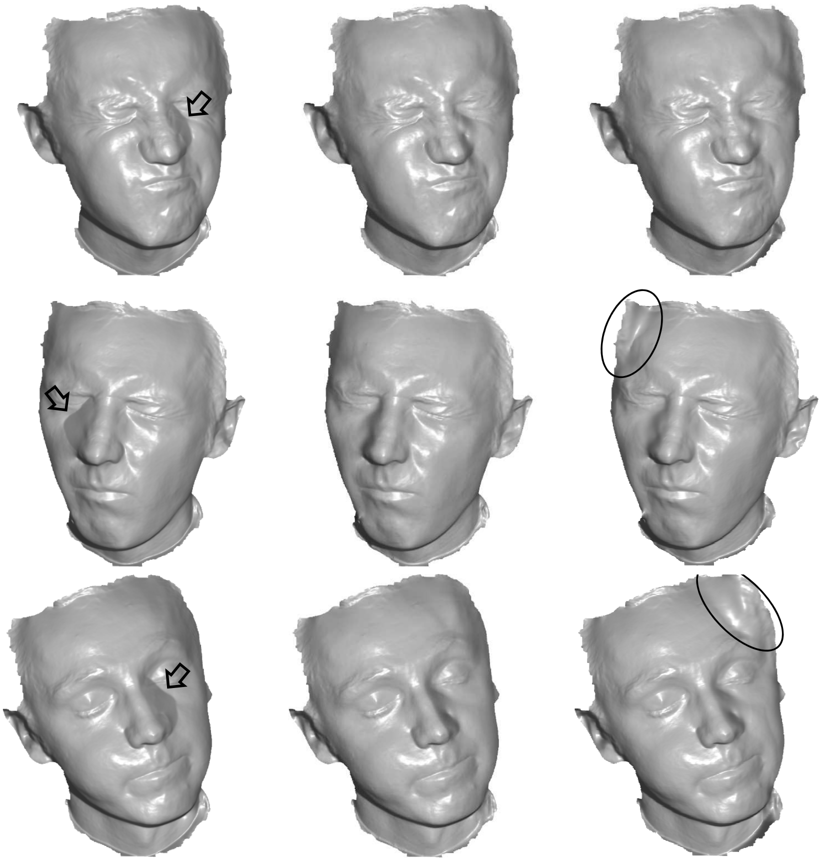


Fig. 6 Face sequence. Three different frames out of a 1000 frame face video sequence. The left column shows the reconstruction when shadows are ignored. Middle and right columns show the corresponding reconstructions after detecting and compensating for the shadow regions using the shading regularisation scheme (middle) and shape regularisation scheme (right). Note the improvement in the regions around the nose reconstruction where strong cast shadows appear (see arrows). Note also how the shape regularisation scheme fails to reconstruct some boundary regions (see circles). This behaviour is further explained in Fig. 7.

what the shape regularisation scheme is missing in order to make the right decision and choose the convex solution.

A second facial performance capture using [19] is shown in figure 8. This time the face is not painted, which implies an assumption of constant albedo chromaticity. In order to cope with shadows, the shading regularisation scheme is used. We observe that, despite the constant albedo deviations, *e.g.* the lips, the system successfully captures fine details such as skin wrinkles.

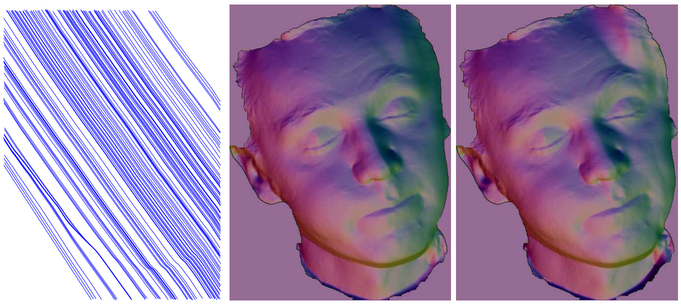


Fig. 7 Failure case of the shape regularisation scheme. The figures correspond to the bottom face in Fig.6. Left shows characteristic curves describing the light occlusion on the right-side of the face. Middle and right show the rendering of the shape under the occluded light using the shading regularisation scheme (middle) and the shape regularisation scheme (right). The failure of the shape regularisation scheme is clearly visible at the top right of the image.

2.7 Related work

The animation and capture of deformations is being explored in many fields, so we provide a general explanation of their relevance in the context of the proposed technique.

Texture Cues

White and Forsyth [48, 49] and Scholz *et al*[41] have presented work on using texture cues to perform the specific task of cloth capture. Their methods are based on printing a special pattern on a piece of cloth and capturing video sequences of that cloth in motion. The estimation of the cloth geometry is based on the observed deformations of the known pattern as well as texture cues extracted from the video sequence. The techniques produce results of very good quality but are ultimately limited by the requirement of printing a special pattern on the cloth which may not be practical for a variety of situations. In the present work, we avoid this requirement while producing detailed results.

Pilet *et al* [38] and Salzmann *et al*[39] proposed a slightly more flexible approach where one uses the pattern already printed in a piece of cloth, by presenting it to the system in a flattened state. Using sparse feature matching the pattern can be detected in each frame of a video sequence. Due to the fact that detection occurs separately in each frame, the method is quite robust to occlusions. However the presented results dealt only with minor non-rigid deformations.

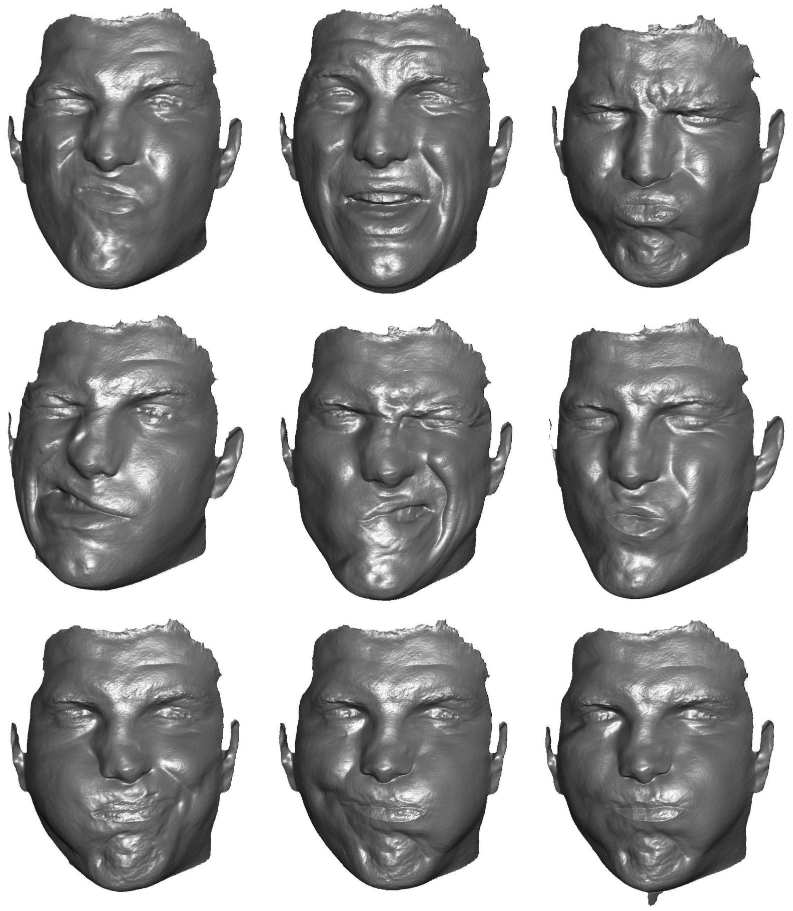


Fig. 8 Face sequence. Acquisition of 3d facial expressions using [19] and the shadow processing technique described in this paper. The shadows are processed with the shading regularisation scheme. The full video sequence has more than a 1000 frames reconstructed.

Photometric Stereo

Photometric stereo [51] is one of the most successful techniques for surface reconstruction from images. It works by observing how changing illumination alters the image intensity of points throughout the object surface. These changes reveal the local surface orientations. This field of local surface orientations can then be integrated into a 3d shape. State of the art photometric-stereo allows uncalibrated light estimation [29, 46] as well as multiple unknown albedos [14, 21]. As mentioned previously, the main difficulty with applying photometric stereo to deforming objects lies in the requirement of changing the light source direction for each captured frame, while the object remains still. This is quite impractical when reconstructing the 3d geometry of a *moving* object. We have shown how multispectral lighting allows one to essentially capture three images (each with a different light direction) in a single snapshot, thus making per-frame photometric reconstruction possible.

Coloured and Structured Lights

The earliest related works are also the most relevant to the method presented in this chapter. The first reference to multispectral light for photometric stereo dates back 20 years to the work of Petrov [37]. Ten years later, Kontsevich *et al* [25] actually demonstrated an algorithm for calibrating unknown color light sources and at the same time computing the surface normals of an object in the scene. They verified the theory on synthetic data and an image of a real egg. We use a simplified approach for calibration and the same orientation-from-colour cue to eventually convert video of un-textured cloth into a single surface with complex changing deformations.

More recently, the parameters needed to simulate realistic cloth dynamics were measured in video by projecting explicitly structured horizontal light stripes onto material samples under static and dynamic conditions [5]. This system measured the edges and silhouette mismatches present in real vs. simulated sequences. Many researchers have utilised structured lighting, and Gu *et al* [16] even used colour, although their method is mostly for storing and manipulating acquired surface models of shading and geometry. Weise *et al*[47] is the current state-of-the-art for structured light and has some advantages in terms of absolute 3d depth, but at the expense of both spatial and temporal sampling, e.g. 17 Hz compared to our 60 Hz (or faster, limited only by the camera used). Zhang *et al*[53] is a nice complete system also with structured lighting that applies to face models and videos. Sand *et al*dispensed with special lighting but leveraged motion capture and automatic silhouettes to deform a human body template [40]. Our technique, on the other hand, expects no prior models of the cloth being reconstructed.

Shadows in Photometric Stereo

One way of characterising photometric stereo methods is based on the number of different lights required and how they cope with highlights or shadows.

A minimum of 3 lights is required to perform photometric stereo with no extra assumptions [51], and only 2 lights with the additional assumption of constant albedo [35]. Whenever more lights are available, the light visibility problem becomes a labelling problem where each point on the surface has to be assigned to the correct set of lights in order to successfully reconstruct the surface.

For objects with constant albedo, [11] used a Rank-2 constraint to detect surfaces illuminated by only 2 lights. In the case of general albedo, every point on the surface has to be visible in at least 3 images. A 4-light photometric stereo setup was proposed in [34], where light occlusion was detected by checking the consistency of all the possible triplets of lights. The work by [52] was able to detect light occlusions in a 4-light setup and simply treat them as outliers. In [2] a similar algorithm to [34] is presented using a 4-light coloured photometric stereo approach.

In the recent work by [6], an iterative MRF formulation is proposed for detecting light occlusion and exploiting it as a surface integration constraint. However, the algorithm also requires a minimum of 4 lights and is targeted for setups with a large number of lights.

3 Multi-view Photometric Stereo

The motivation for the method presented in this section is digital archiving of 3d objects, a key area of interest in cultural heritage preservation. While laser range scanning is one of the most popular techniques, it has a number of drawbacks, namely the need for specialised, expensive hardware and also the requirement of exclusive access to an object for significant periods of time. Also, for a large class of shiny objects such as porcelain or glazed ceramics, 3d scanning with lasers is challenging [27]. Recovering 3d shape from photographic images is an efficient, cost effective way to generate accurate 3d scans of objects.

Several solutions have been proposed for this long studied problem. When the object is well textured its shape can be obtained by densely matching pixel locations across multiple images and triangulating (see previous chapter or [42] for a recent review), however the results typically exhibit high frequency noise.

For non textured objects photometric stereo is a well established alternative that can provide very detailed reconstructions. One of the biggest drawbacks of photometric stereo methods is the fact that they can only provide single-viewpoint, 2.5D depth-map reconstructions.

In this section we describe an elegant and practical method for acquiring a *complete* and *accurate* 3d model from a number of images taken around the object, captured under changing light conditions (see Fig. 9). The changing (but otherwise unknown) illumination conditions uncover the fine geometric detail of the object surface which is obtained by a generalised photometric stereo scheme.

The object's reflectance is assumed to follow Lambert's law, *i.e.* points on the surface keep their appearance constant irrespective of viewpoint. The method can however tolerate isolated specular highlights, typically observed in glazed surfaces such as porcelain. We also assume that a single, distant light-source illuminates the object and that it can be changed arbitrarily between image captures. Finally, it is assumed that the object can be segmented from the background and silhouettes extracted automatically.

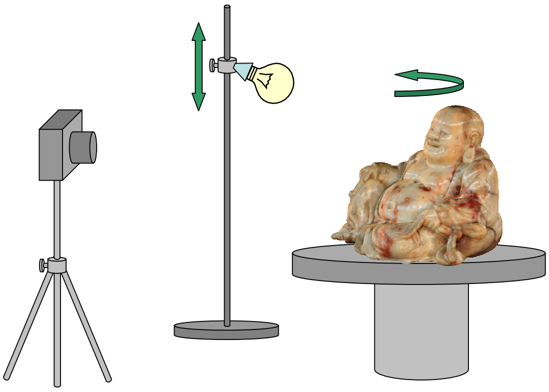


Fig. 9 Our acquisition setup. The object is rotated on a turntable in front of a camera and a point light-source. A sequence of images are captured while the light-source changes position between consecutive frames. No knowledge of the camera or light-source positions is assumed.

3.1 Related work

The method presented here draws inspiration from the recent work of [29] where the authors explore the possibility of using photometric stereo with images from multiple views, when correspondence between views is not initially known. Picking an arbitrary viewpoint as a reference image, a depth-map with respect to that view serves as the source of approximate correspondences between frames. This depth-map is initialised from a Delaunay triangulation of sparse 3d features located on the surface. Using this depth-map, their algorithm performs a photometric stereo computation obtaining normal directions for each depth-map location. When these normals are integrated, the resulting depth-map is closer to the true surface than the original. The paper presents high quality reconstructions and gives a theoretical argument justifying the convergence of the scheme. The method however relies on the existence of distinct features on the object surface which are tracked to obtain camera motion and initialise the depth-map. In the class of textureless objects we are considering, it may be impossible to locate such surface features and indeed our method has no such requirement. Also the surface representation is still depth-map based and consequently the models produced are 2.5D.

A similar approach of extending photometric stereo to multiple views and more complex BRDFs was presented in [36] with the limitation of almost planar 2.5D reconstructed surfaces. Our method is based on the same fundamental principle of bootstrapping photometric stereo with approximate correspondences, but we use a general volumetric framework which allows complete 3d reconstructions from multiple views.

Quite related to this idea is the work of [3] and [33] where photometric stereo information is combined with 3d range scan data. In [3] the photometric information is simply used as a normal map texture for visualisation purposes. In [33], a very good initial approximation to the object surface is obtained using range scanning technology, which however is shown to suffer from high-frequency noise. By applying a fully calibrated 2.5D photometric stereo technique, normal maps are estimated which are then integrated to produce an improved, almost noiseless surface geometry. Our acquisition technique is different from [33] in the following respects: (1) we only use standard photographic images and simple light sources, (2) our method is fully uncalibrated- all necessary information is extracted from the object's contours and (3) we completely avoid the time consuming and error prone process of merging 2.5D range scans.

The use of the silhouette cue is inspired by the work of [45] where a scheme for the recovery of illumination information, surface reflectance and geometry is described. The algorithm described makes use of frontier points, a geometrical feature of the object obtained by the silhouettes. Frontier points are points of the visual hull where two contour generators intersect and hence are guaranteed to be on the object surface. Furthermore the local surface orientation is known at these points, which makes them suitable for various photometric computations such as extraction of reflectance and illumination information. Our method generalises the idea by examining a much richer superset of frontier points which is the set of contour generator points. We overcome the difficulty of localising contour generators by a robust random sampling strategy. The price we pay is that a considerably simpler reflectance model must be used.

Although solving a different type of problem, the work of [23] is also highly related mainly because the class of objects addressed is similar to ours. While the energy term defined and optimised in their paper bears strong similarity to ours, their reconstruction setup keeps the lights fixed with respect to the object so in fact an entirely different problem is solved and hence a performance comparison between the two techniques is difficult. However the results presented in [23] at first glance seem to be lacking in detail especially in concavities, while our technique considerably improves on the visual hull. Finally, there is a growing volume of work on using specularities for calibrating photometric stereo (see [10] for a detailed literature survey). This is an example of a different cue used for performing uncalibrated photometric stereo on objects of the same class as the one considered here. However methods proposed have so far only been concerned with the fixed view case.

3.2 Algorithm

The method presented here reconstructs the complete geometry of 3d objects by exploiting the powerful silhouette and shading cues. We modify classic photometric stereo and cast it in a multi-view framework where the camera is allowed to circumnavigate the object and illumination is allowed to vary. Firstly, the object's silhouettes are used to recover camera motion using the technique presented in [18], and via a novel robust estimation scheme they allow us to accurately estimate the light directions and intensities in every image.

Secondly, the object surface, which is parametrised by a mesh and initialised from the visual hull, is evolved until its predicted appearance matches the captured images. The advantages of our approach are the following:

- It is fully uncalibrated: no light or camera pose calibration object needs to be present in the scene. Both camera pose and illumination are estimated from the object's silhouettes.
- The full 3d geometry of a complex, textureless multi-albedo object is accurately recovered, something not previously possible by any other method.
- It is practical and efficient as evidenced by our simple acquisition setup.

3.2.1 Robust estimation of light-sources from the visual hull

For an image of a lambertian object with varying albedo, under a single distant light source, and assuming no self-occlusion, each surface point projects to a point of intensity given by:

$$c = \mathbf{l}^T \rho \mathbf{n}, \tag{15}$$

where $\mathbf{l}$ is a 3d vector directed towards the light-source and scaled by the light-source intensity, $\mathbf{n}$ is the surface unit normal at the object location and ρ is the albedo at that location. Equation (15) provides a single constraint on the three coordinates of the product $\rho\mathbf{l}$. Then, given three points $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$ with an unknown but *equal* albedo ρ, their normals (non co-planar) $\mathbf{n_1}, \mathbf{n_2}, \mathbf{n_3}$, and the corresponding three image intensities i_1, i_2, i_3, we can construct three such equations that can uniquely determine $\rho\mathbf{l}$ as

$$\rho\mathbf{l} = [\mathbf{n_1}\,\mathbf{n_2}\,\mathbf{n_3}]^{-1} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}. \tag{16}$$

For multiple images, these same three points can provide the light directions and intensities in each image up to a global unknown scale factor ρ. The problem is then how to obtain three such points.

Our approach is to use the powerful silhouette cue. The observation on which this is based is the following: when the images have been calibrated for camera motion, the object's silhouettes allow the construction of the *visual hull* [26], which is defined as the maximal volume that projects inside the silhouettes (see Fig. 10). A fundamental property of the visual hull is that its surface coincides with the real surface of the object along a set of 3d curves, one for each silhouette, known as *contour generators* [8]. Furthermore, for all points on those curves, the surface orientation of the visual hull surface is equal to the orientation of the object surface. Therefore if we could detect points on the visual hull that belong to contour generators and have equal albedo, we could use their surface normal directions and projected intensities to estimate lighting. Unfortunately contour generator points with equal albedo cannot be directly identified within the set of all points of the visual hull. Light estimation however can be viewed as robust model fitting where the inliers are the contour generator points of some constant albedo and the outliers are the

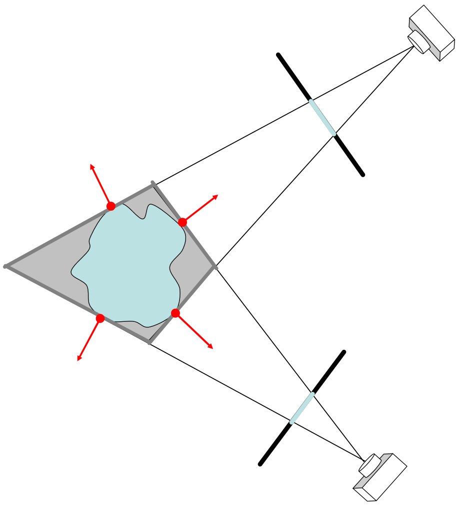


Fig. 10 The visual hull for light estimation. The figure shows a 2d example of an object which is photographed from two viewpoints. The visual hull (gray quadrilateral) is the largest volume that projects inside the silhouettes of the object. While the surface of the visual hull is generally quite far from the true object surface, there is a set of points where the two surfaces are tangent and moreover, share the same local orientation (these points are denoted here with the four dots and arrows). In the full 3d case, three points with their surface normals, are enough to fix an illumination hypothesis, against which all other points can be tested for agreement. This suggests a robust random sampling scheme, described in the main text, via which the correct illumination can be obtained.

rest of the visual hull points. The albedo of the inliers will be the *dominant* albedo, *i.e.* the colour of the majority of the contour generator points. One can expect that the outliers do not generate consensus in favour of any particular illumination model while the inliers do so in favour of the correct model. This observation motivates us to use a robust RANSAC scheme [13] to separate inliers from outliers and estimate illumination direction and intensity. The scheme can be summarised as follows:

1. Pick three points on the visual hull and from their image intensities and normals estimate an illumination hypothesis for $\rho\mathbf{l}$.
2. Every point on the visual hull $\mathbf{x_m}$ will now vote for this hypothesis *if* its predicted image intensity is within a given threshold τ of the observed image intensity c_m, *i.e.*
$$\left|\rho\mathbf{l}^T\cdot\mathbf{n_m}-c_m\right|<\tau, \tag{17}$$
where τ allows for quantisation errors, image noise, etc.
3. Repeat 1 and 2 a set number of times always keeping the illumination hypothesis with the largest number of votes.

The shape of the actual function being optimised by the RANSAC scheme described above was explored graphically for a porcelain object in Fig. 11. The num-

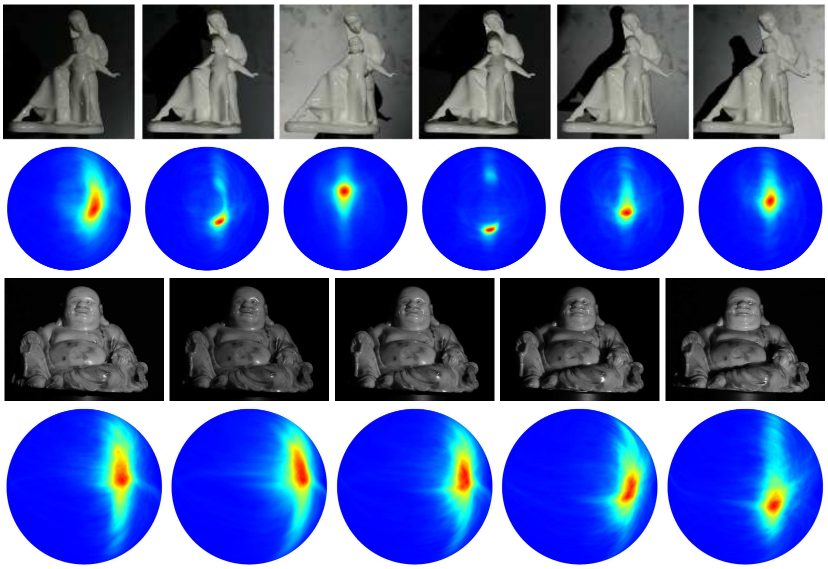


Fig. 11 Shape of illumination consensus. For different illumination configurations we have plotted the consensus as a function of light direction. For each direction consensus has been maximised with respect to light intensity. Red values denote big consensus. The shape of the maxima of this cost function as well as the lack of local optima implies a stable optimisation problem. Top: 6 different illuminations of a single albedo object. Bottom: 4 different illuminations of a multi-albedo object. Although the presence of multiple albedos degrades the quality of the light estimation (the peak is broader), it is still a clear single optimum.

ber of points voting for a light direction (maximised with respect to light intensity) was plotted as a 2d function of latitude and longitude of the light direction. These graphical representations, obtained for six different illuminations, show the lack of local optima and the presence of clearly defined maxima.

This simple method can also be extended in the case where the illumination is kept fixed with respect to the camera for K frames. This corresponds to K illumination vectors $R_1\mathbf{l},\ldots,R_K\mathbf{l}$ where R_k are 3×3 rotation matrices that rotate the fixed illumination vector $\mathbf{l}$ with respect to the object. In that case a point on the visual hull $\mathbf{x_m}$ with normal $\mathbf{n_m}$ will vote for $\mathbf{l}$ if it is visible in the k-th image where its intensity is $c_{m,k}$ and

$$\left|\rho(R_k\mathbf{l})^T \cdot \mathbf{n_m} - c_{m,k}\right| < \tau. \tag{18}$$

A point is allowed to vote more than once if it is visible in more than one image.

Even though in theory the single image case suffices for independently recovering illumination in each image, in our acquisition setup light can be kept fixed over more than one frame. This allows us to use the extended scheme in order to further improve our estimates. A performance comparison between the single view and the multiple view case is provided through simulations with synthetic data in the experiments section.

An interesting and very useful byproduct of the robust RANSAC scheme is that any deviations from our assumptions of a Lambertian surface of uniform albedo are rejected as outliers. This provides the light estimation algorithm with a degree of tolerance to sources of error such as highlights or local albedo variations. The next section describes the second part of the algorithm which uses the estimated illumination directions and intensities to recover the object surface.

3.2.2 Fusing multiple views

Having estimated the distant light-source directions and intensities for each image our goal is to find a closed 3d surface that is photometrically consistent with the images and the estimated illumination, *i.e.* its predicted appearance by the lambertian model and the estimated illumination matches the images captured. To achieve this we use an optimisation approach where a cost function penalising the discrepancy between images and predicted appearance is minimised.

Our algorithm optimises a surface S that is represented as a mesh with vertices $\mathbf{x_1} \dots \mathbf{x_M}$, triangular faces $f = 1 \dots F$ and corresponding albedo $\rho_1, \dots, \rho_F$. We denote by $\mathbf{n_f}$ and A_f the mesh normal and the surface area at face f. Also let $c_{f,k}$ be the intensity of face f on image k and let the set $\mathcal{V}_f$ be the set of images (subset of $\{1, \dots, K\}$) from which face f is visible. The light direction and light intensity of the k-th image will be denoted by $\mathbf{l_k}$.

We use a scheme similar to the ones used in [23, 46] where the authors introduce a decoupling between the mesh normals $\mathbf{n_1} \dots \mathbf{n_F}$, and the direction vectors used in the Lambertian model equation. We call these new direction vectors $\mathbf{v_1} \dots \mathbf{v_F}$ *photometric normals*, and they are independent of the mesh normals. The minimisation cost is then composed of two terms, where the first term E_v links the photometric normals to the observed image intensities:

$$E_v\left(\mathbf{v_{1,\dots,F}}, \rho_{1,\dots,F}; \mathbf{x_{1,\dots,M}}\right) = \sum_{f=1}^{F} \sum_{k \in \mathcal{V}_f} \left(\mathbf{l_k}^T \rho_f \mathbf{v_f} - c_{f,k}\right)^2, \tag{19}$$

and the second term E_m brings the mesh normals close to the photometric normals through the following equation:

$$E_m\left(\mathbf{x_{1,\dots,M}}; \mathbf{v_{1,\dots,F}}\right) = \sum_{f=1}^{F} \|\mathbf{n_f} - \mathbf{v_f}\|^2 A_f. \tag{20}$$

This decoupled energy function is optimised by iterating the following two steps:

1. **Photometric normal optimisation.** The vertex locations are kept fixed while E_v is optimised with respect to the photometric normals and albedos. This is achieved by solving the following independent minimisation problems for each

face f:

$$\mathbf{v_f}, \rho_f = \arg\min_{\mathbf{v},\rho} \sum_{k \in \mathcal{V}_f} \left(\mathbf{l_k}^T \rho \mathbf{v} - c_{f,k} \right)^2 \text{ s.t. } ||\mathbf{v}|| = 1. \tag{21}$$

2. **Vertex optimisation.**The photometric normals are kept fixed while E_m is optimised with respect to the vertex locations using gradient descent.

These two steps are interleaved until convergence which takes about 20 steps for the sequences we experimented with. Typically each integration phase takes about 100 gradient descent iterations. Note that for the first step described above, *i.e.* evolving the mesh until the surface normals converge to some set of *target* orientations, a variety of solutions is possible. A slightly different solution to the same geometric optimisation problem has recently been proposed in [33], where the target orientations are assigned to each vertex, rather than each face as we do here. That formulation lends itself to a closed-form solution with respect to the position of a single vertex. An iteration of these local vertex displacements yields the desired convergence. As both formulations offer similar performance, the choice between them should be made depending on whether the target orientations are given on a per vertex or per facet basis.

The visibility map $\mathcal{V}_f$ is a set of images in which we can measure the intensity of face f. It excludes images in which face f is occluded using the current surface estimate as the occluding volume as well as images where face f lies in shadow. Shadows are detected by a simple thresholding mechanism, *i.e.* face f is assumed to be in shadow in image k if $c_{f,k} < \tau_{shadow}$ where τ_{shadow} is a sufficiently low intensity threshold. Due to the inclusion of a significant number of viewpoints in $\mathcal{V}_f$, (normally at least 4) the system is quite robust to the choice of τ_{shadow}. For all the experiments presented here, the value $\tau_{shadow} = 5$ was used (for intensities in the range 0-255). As for the highlights, we also define a threshold $\tau_{highlight}$ such as a face f is assumed to be on a highlight in image k if $c_{f,k} > \tau_{highlight}$. In order to compute $\tau_{highlight}$ need to distinguish between single albedo objects and multi-albedo objects. Single albedo objects are easily handled since the light calibration step gives us the light intensity. Hence, under the Lambertian assumption, no point on the surface can produce an intensity higher than the light intensity, *i.e.* $\tau_{highlight} = ||\rho \mathbf{l}||$. In the multi-albedo case ρ can also vary, and it is likely that the albedo picked by the robust light estimation algorithm is not the brightest one present on the object. As a result, we prefer to use a global threshold to segment the highlights on the images. It is worth noting that this approach works for the porcelain objects because highlights are very strong and localised, so just a simple sensor saturation test is enough to find them, *i.e.* $\tau_{highlight} = 254$.

3.3 Experiments

The setup used to acquire the 3d model of the object is quite simple (see Fig. 9). It consists of a turntable, onto which the object is mounted, a 60W halogen lamp and

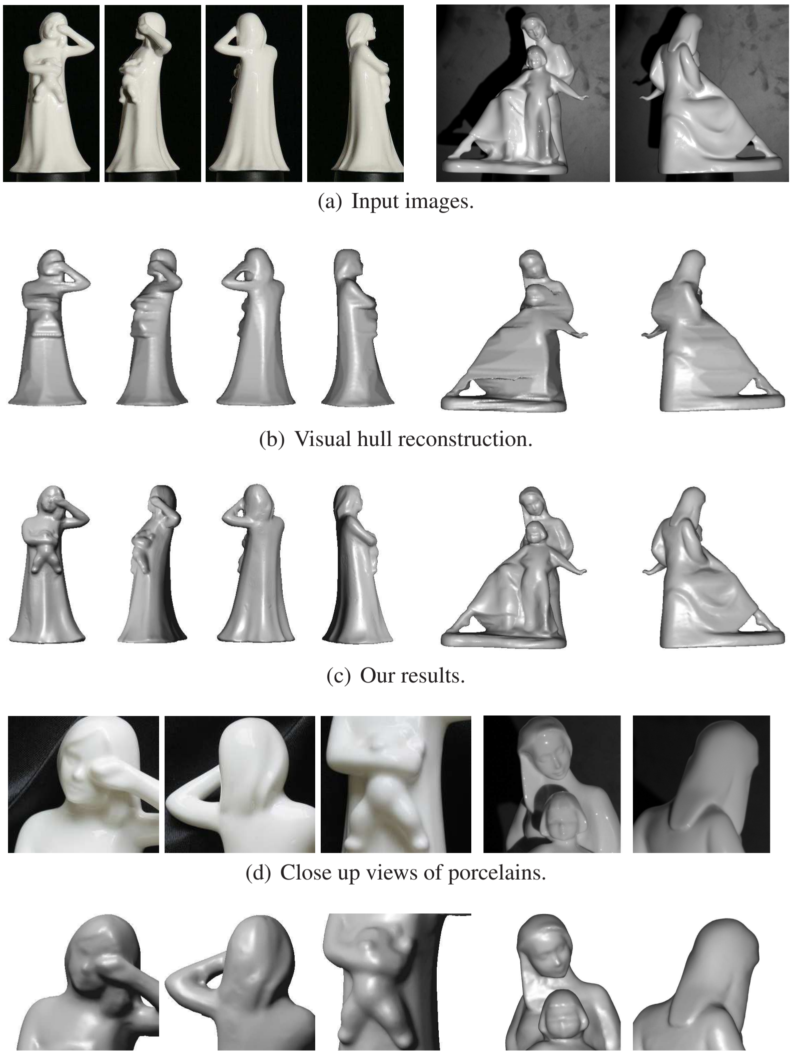
```
Capture images of object.
Extract silhouettes.
Recover camera motion and compute visual hull.
Estimate light directions and intensities in every image (Section 3.2.1).
Initialise a mesh with vertices $\mathbf{x_1} \dots \mathbf{x_M}$ and faces $f = 1 \dots F$ to the object's visual hull.
while mesh-not-converged do
    Optimise $E_v$ with respect to $\mathbf{v_1} \dots \mathbf{v_F}$ (19).
    Optimise $E_m$ with respect to $\mathbf{x_1} \dots \mathbf{x_M}$ (20).
end while
```

Fig. 12 The multi-view reconstruction algorithm.

a digital camera. The object rotates on the turntable and 36 images (*i.e.* a constant angle step of 10 degrees) of the object are captured by the camera while the position of the lamp is changed. In our experiments we have used three different light positions which means that the position of the lamp was changed after twelve, and again after twenty-four frames. The distant light source assumptions are satisfied if an object of 15cm extent is placed 3-4m away from the light.

The algorithm was tested on five challenging shiny objects, two porcelain figurines shown in Fig. 13, two fine relief Chinese Qing-dynasty porcelain vases shown in Fig. 14, and one textured Jade Buddha figurine in Fig. 15. Thirty-six 3456×2304 images of each of the objects were captured under three different illuminations. The object silhouettes were extracted by intensity thresholding and were used to estimate camera motion and construct the visual hull (second row of Fig. 13). The visual hull was processed by the robust light estimation scheme of Section 3.2.1 to recover the distance light-source directions and intensities in each image. The photometric stereo scheme of section 3.2.2 was then applied. The results in Fig. 14 show reconstructions of porcelain vases with very fine relief. The reconstructed relief (especially for the vase on the right) is less than a millimetre while their height is approximately 15-20 cm. Figure 15 shows a detailed reconstruction of a Buddha figurine made of polished Jade. This object is actually textured, which implies classic stereo algorithms could be applied. Using the camera motion information and the captured images, a state-of-the-art multi-view stereo algorithm [17] was executed. The results are shown in the second row of Figure 15. It is evident that, while the low frequency component of the geometry of the figurine is correctly recovered, the high frequency detail obtained by [17] is noisy. The reconstructed model appears bumpy even though the actual object is quite smooth. Our results do not exhibit surface noise while capturing very fine details such as surface cracks.

To quantitatively analyse the performance of the multi-view photometric stereo scheme presented here with ground truth, an experiment on a synthetic scene was performed (Fig. 16). A 3d model of a sculpture (digitised via a different technique) was rendered from 36 viewpoints with uniform albedo and using the Lambertian reflectance model. The 36 frames were split into three sets of 12 and within each set the single distant illumination source was held constant. Silhouettes were extracted from the images and the visual hull was constructed. This was then used to estimate the illumination direction and intensity as described in Section 3.2.1. In 1000 runs of



(a) Input images.

(b) Visual hull reconstruction.

(c) Our results.

(d) Close up views of porcelains.

(e) Close up views of reconstructed models.

Fig. 13 Reconstructing porcelain figurines. Two porcelain figurines reconstructed from a sequence of 36 images each (some of the input images are shown in (a)). The object moves in front of the camera and illumination (a 60W halogen lamp) changes direction twice during the image capture process. (b) shows the results of a visual hull reconstruction while (c) shows the results of our algorithm. (d) and (e) show detailed views of the figurines and the reconstructed models respectively.

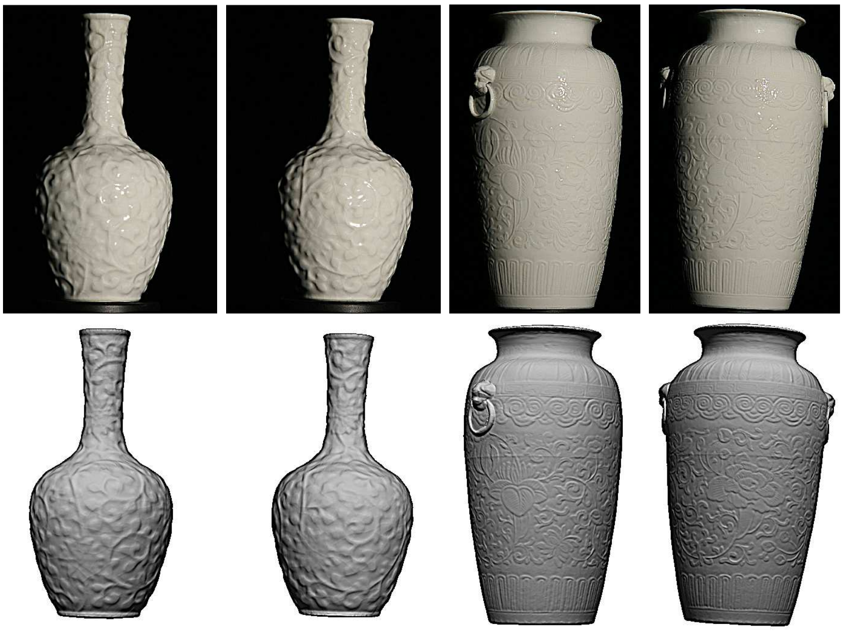


Fig. 14 Reconstructing Chinese Qing-dynasty porcelain vases. Top: sample of input images. Bottom: proposed method. The resulting surface captures all the fine details present in the images, even in the presence of strong highlights.

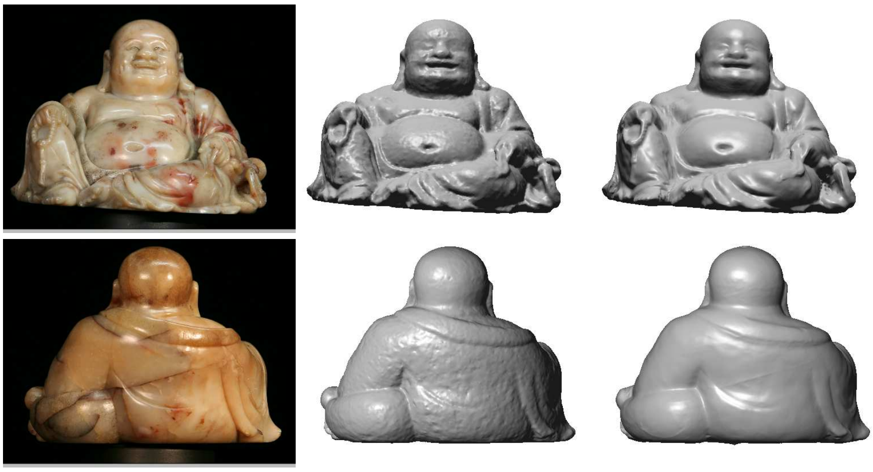


Fig. 15 Reconstructing coloured jade. Left: Two input images. Middle: model obtained by multi-view stereo method from [17]. Right: proposed method. The resulting surface is filtered from noise while new high frequency geometry is revealed (note the reconstructed surface cracks in the middle of the figurine's back).

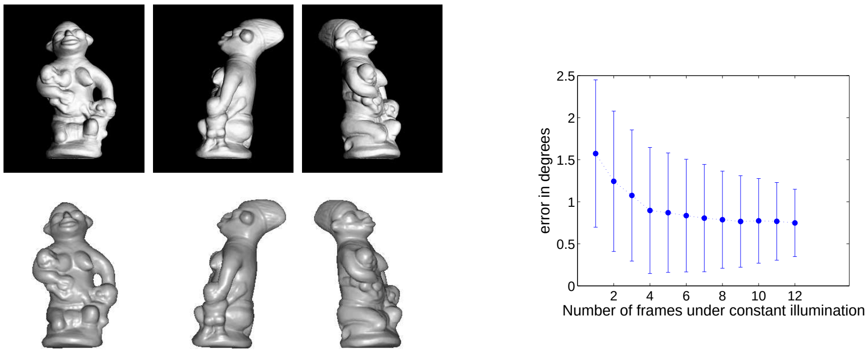


Fig. 16 Synthetic evaluation. Left: the accuracy of the algorithm was evaluated using an image sequence synthetically generated from a 3d computer model of a sculpture. This allowed us to compare the quality of the reconstructed model against the original 3d model as well as measure the accuracy of the light estimation. The figure shows the reconstruction results obtained, below the images of the synthetic object. The mean distance of all points of the reconstructed model from the ground truth was found to be about 0.5mm if the bounding volume's diagonal is 1m. Right: The figure shows the effect of varying the length of the frame subsequences that have constant light. The angle between the recovered light direction and ground truth has been measured for 1000 runs of the RANSAC scheme for each number of frames under constant lighting. With just a single frame per illumination the algorithm achieves a mean error of 1.57 degrees with a standard deviation of 0.88 degrees. With 12 frames sharing the same illumination the mean error drops to 0.75 degrees with a standard deviation of 0.41 degrees.

the illumination estimation method for the synthetic scene, the mean light direction estimate was 0.75 degrees away from the true direction with a standard deviation of 0.41 degrees. The model obtained by our algorithm was compared to the ground truth surface by measuring the distance of each point on our model from the closest point in the ground truth model. This distance was found to be about 0.5mm when the length of the biggest diagonal of the bounding box volume was defined to be 1m. Even though this result was obtained from perfect noiseless images it is quite significant since it implies that any loss of accuracy can only be attributed to the violations of our assumptions rather than the optimisation methods themselves. Many traditional multi-view stereo methods would not be able to achieve this due to the strong regularisation that must be imposed on the surface. By contrast our method requires no regularisation when faced with perfect noiseless images.

Finally, we investigated the effect of the number of frames during which illumination is held constant with respect to the camera frame. Our algorithm can in theory obtain the illumination direction and intensity in every image independently. However keeping the lighting fixed over two or more frames, and supplying that knowledge to the algorithm can significantly improve estimates. The next experiment was designed to test this improvement by performing a light estimation over K images where the light has been kept fixed with respect to the camera. The results are plotted in Figure 16 right and show the improvement of the accuracy of the recovered lighting directions as K increases from 1 to 12. The metric used was the angle between the ground truth light direction and the estimated light direction over 1000 runs of the robust estimation scheme. For $K = 1$ the algorithm achieves a mean error of 1.57 degrees with a standard deviation of 0.88 while for $K = 12$ it achieves 0.75 degrees with a standard deviation of 0.41 degrees. The decision for selecting a value for K should be a consideration of the tradeoff between practicality and maximising the total number of different illuminations in the sequence which is M/K where M is the total number of frames.

References

1. Amit, G.F., Agrawal, A., Raskar, R.: What is the range of surface reconstructions from a. In: Proc. 9^{th} Europ. Conf. on Computer Vision (ECCV), pp. 578–591. Springer (2006)
2. Barsky, S., Petrou, M.: The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1239–1252 (2003)
3. Bernardini, F., Rushmeier, H., Martin, I., Mittleman, J., Taubin, G.: Building a digital model of michelangelo's florentine pieta. IEEE Computer Graphics and Applications **22**(1), 59–67 (2002)
4. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: SIGGRAPH '00, pp. 417–424. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (2000). DOI http://doi.acm.org/10.1145/344779.344972
5. Bhat, K.S., Twigg, C.D., Hodgins, J.K., Khosla, P.K., Popović, Z., Seitz, S.M.: Estimating cloth simulation parameters from video. In: SCA '03: Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer animation, pp. 37–51 (2003)
6. Chandraker, M., Agarwal, S., Kriegman, D.: Shadowcuts: Photometric stereo with shadows. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2007)
7. Chen, H., Belhumeur, P., Jacobs, D.: In search of illumination invariants. In: Proc. IEEE Conf. CVPR, vol. 2, pp. 254–261 (2000)
8. Cipolla, R., Giblin, P.: Visual Motion of curves and surfaces. Cambridge University Press (1999)
9. Davis, T.A.: Algorithm 832: Umfpack, an unsymmetric-pattern multifrontal method. ACM Transactions on Mathematical Software **30**(2), 196–199 (2004)
10. Dbrohlav, O., Chandler, M.: Can two specular pixels calibrate photometric stereo ? In: Proc. 10^{th} Intl. Conf. on Computer Vision (ICCV) (2005)
11. Drew, M.: Reduction of rank-reduced orientation-from-color problem with many unknown lights to two-image known-illuminant photometric stereo. In: IEEE International Symposium on Computer Vision, pp. 419–424 (1995)
12. Fan, J., Wolff, L.B.: Surface curvature and shape reconstruction from unknown multiple illumination and integrability. Comput. Vis. Image Underst. **65**(2), 347–359 (1997). DOI http://dx.doi.org/10.1006/cviu.1996.0581
13. Fischler, M.A., Bolles, R.C.: Ransac, random sampling consensus: a paradigm for model fitting with applications to image analysis and autoomated cartography. Comm. ACM **26**, 381–395 (1981)
14. Goldman, D.B., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and spatially-varying BRDFs from photometric stereo. In: ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), vol. 1, pp. 341–348 (2005)
15. Greig, D., Porteous, B., Seheult., A.: Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society **51**(2), 271–279 (1989)
16. Gu, X., Zhang, S., Huang, P., Zhang, L., Yau, S.T., Martin, R.: Holoimages. In: SPM '06: Proceedings of the 2006 ACM Symposium on Solid and Physical Modeling, pp. 129–138. ACM Press (2006)

17. Hernández, C., Schmitt, F.: Silhouette and stereo fusion for 3d object modeling. Computer Vision and Image Understanding **96**(3), 367–392 (2004)
18. Hernández, C., Schmitt, F., Cipolla, R.: Silhouette coherence for camera calibration under circular motion. IEEE Trans. Pattern Anal. Mach. Intell. **29**(2), 343–349 (2007)
19. Hernández, C., Vogiatzis, G., Brostow, G., Stenger, B., Cipolla, R.: Non-rigid photometric stereo with colored lights. In: Proc. 11^{th} Intl. Conf. on Computer Vision (ICCV) (2007)
20. Hertzmann, A., Seitz, S.: Shape and materials by example: a photometric stereo approach. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. I: 533–540 (2003)
21. Hertzmann, A., Seitz, S.: Shape reconstruction with general, varying brdfs. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1254–1264 (2005)
22. Horn, B.K.P.: Robot vision. MIT press (1986)
23. Jin, H., Cremers, D., Yezzi, A., Soatto, S.: Shedding light in stereoscopic segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 36–42 (2004)
24. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10), 1568–1583 (2006). DOI http://dx.doi.org/10.1109/TPAMI.2006.200
25. Kontsevich, L., Petrov, A., Vergelskaya, I.: Reconstruction of shape from shading in color images. J. Opt. Soc. Am. A **11**(3), 1047–1052 (1994)
26. Laurentini, A.: The visual hull concept for silhouette-based image understanding. IEEE Trans. Pattern Anal. Mach. Intell. **16**(2) (1994)
27. Levoy, M.: Why is 3d scanning hard? Invited address at 3D Processing, Visualization, Transmission, Padua, Italy (2002)
28. Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., Shade, J., Fulk, D.: The digital michelangelo project: 3d scanning of large statues. In: Proc. of the ACM SIGGRAPH, p. 1522 (2000)
29. Lim, J., Ho, J., Yang, M.H., Kriegman, D.: Passive photometric stereo from motion. In: ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision, pp. 1635–1642 (2005)
30. Lin, S., Lee, S.: Estimation of diffuse and specular appearance. In: Proc. 7^{th} Intl. Conf. on Computer Vision (ICCV), vol. 2, pp. 855–860 (1999)
31. Malzbender, T., B. Wilburn, D.G., Ambrisco, B.: Surface enhancement using real-time photometric stereo and reflectance transformation. In: Eurographics Symposium on Rendering 2006. Nicosia, Cyprus (2006)
32. Nayar, S., Ikeuchi, K., Kanade, T.: Surface reflection: physical and geometrical perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **13(7)**, 611–634 (1991)
33. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3d geometry. In: Proc. of the ACM SIGGRAPH, pp. 536–543 (2005)
34. North Coleman Jr., E., Jain, R.: Shape recovery, chap. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry, pp. 180–199. Jones and Bartlett Publishers, Inc., , USA (1992)
35. Onn, R., Bruckstein, A.: Integrability disambiguates surface recovery in two-image photometric stereo. Intl. Journal of Computer Vision **5**(1), 105–113 (1990)
36. Paterson, J., Claus, D., Fitzgibbon, A.: Brdf and geometry capture from extended inhomogeneous samples using flash photography. In: Proc. of Eurographics 2005 (2005)
37. Petrov, A.: Light, color and shape. Cognitive Processes and their Simulation (in Russian) pp. 350–358 (1987)
38. Pilet, J., Lepetit, V., Fua, P.: Real-time non-rigid surface detection. In: Conference on Computer Vision and Pattern Recognition, San Diego, CA (2005)
39. Salzmann, M., Ilic, S., Fua, P.: Physically valid shape parameterization for monocular 3-d deformable surface tracking. In: British Machine Vision Conference (2005)
40. Sand, P., McMillan, L., Popović, J.: Continuous capture of skin deformation. ACM Trans. Graph. **22**(3), 578–586 (2003)

41. Scholz, V., Stich, T., Keckeisen, M., Wacker, M., Magnor, M.: Garment motion capture using color-coded patterns. Computer Graphics Forum (Proc. Eurographics EG'05) **24**(3), 439–448 (2005)
42. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 519–528 (2006)
43. Tankus, A., Kiryati, N.: Photometric stereo under perspective projection. In: Proc. 10^{th} Intl. Conf. on Computer Vision (ICCV), pp. 611–616. IEEE Computer Society, Washington, DC, USA (2005). DOI http://dx.doi.org/10.1109/ICCV.2005.190
44. Treuille, A., Hertzmann, A., Seitz, S.: Example-based stereo with general brdfs. In: Proc. 8^{th} Europ. Conf. on Computer Vision (ECCV) (2004)
45. Vogiatzis, G., Favaro, P., Cipolla, R.: Using frontier points to recover shape, reflectance and illumination. In: Proc. 10^{th} Intl. Conf. on Computer Vision (ICCV), pp. 228–235 (2005)
46. Vogiatzis, G., Hernández, C., Cipolla, R.: Reconstruction in the round using photometric normals and silhouettes. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1847–1854 (2006)
47. Weise, T., Leibe, B., Gool, L.V.: Fast 3d scanning with automatic motion compensation. In: CVPR '07: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2007)
48. White, R., Forsyth, D.: Combining cues: Shape from shading and texture. In: Computer Vision and Pattern Recognition, vol. 2, pp. 1809–1816 (2006)
49. White, R., Forsyth, D.: Retexturing single views using texture and shading. In: European Conference on Computer Vision, vol. LNCS 3954, pp. 70–81. Springer (2006)
50. Wolff, L.B., Angelopoulou, E.: 3d stereo using photometric ratios. In: ECCV '94: Proceedings of the Third European Conference-Volume II on Computer Vision, pp. 247–258. Springer-Verlag, London, UK (1994)
51. Woodham, R.: Photometric method for determining surface orientation from multiple images. Optical Engineering **19**(1), 139–144 (1980)
52. Yuille, A., Snow, D.: Shape and albedo from multiple images using integrability. In: CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), p. 158. IEEE Computer Society, Washington, DC, USA (1997)
53. Zhang, L., Snavely, N., Curless, B., Seitz, S.M.: Spacetime faces: High-resolution capture for modeling and animation. In: ACM Annual Conference on Computer Graphics, pp. 548–558 (2004)