Single-view Perspective Shape-from-texture with Focal Length Estimation: A Piecewise Affine Approach

Toby Collins, Jean-Denis Durou, Adrien Bartoli, Pierre Gurdjos

Single-view Perspective Shape-from-Texture with Focal Length Estimation: A Piecewise Affine Approach

Toby COLLINS
Université d'Auvergne, Clermont-Ferrand
toby.collins@gmail.com

Jean-Denis DUROU
IRIT, Université Paul Sabatier, Toulouse
CMLA, ENS Cachan, CNRS, UniverSud, Cachan
durou@irit.fr

Pierre GURDJOS
IRIT, ENSEEIHT, Toulouse
pgurdjos@enseeiht.fr

Adrien BARTOLI
Université d'Auvergne, Clermont-Ferrand
adrien.bartoli@gmail.com

Abstract

We present a new formulation to the well known problem of shape-from-texture from a single image by casting the task as a multi-plane based camera pose estimation problem. Our first contribution is methodological: we show that by using a piecewise affine model, instead of a perspective one, we can avoid the numerical instabilities in the estimation of the surface pose compared with the full-perspective model, yet retaining high accuracy. Our second contribution is to show that the information provided by a smooth textured surface makes it possible to perform shape-from-texture and camera focal length calibration jointly. This advances state-of-the-art where a calibrated camera is nearly always assumed in order to compute 3D shape from a single image. We validate both these contributions on simulated and real image data.

1. Introduction

In this paper we focus on the open problem of computing the 3D shape of a surface using the Shape-From-Texture (SFT) cue. In the classical SFT setting the following assumptions are made:

- A single view of the surface is provided on which multiple occurrences of one or more patterns are printed in distinct, but not necessarily regularly arranged positions (Fig. 1).
- The fronto-parallel appearance of the pattern (i.e. a template) is known *a priori*.
- The pattern is *small enough* such that each occurrence can be approximated by a planar patch.

Of course, the informative cue is the warp induced by the projection onto the image plane. The image of a patch is called a *texton*.

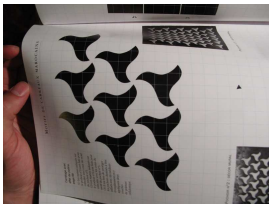


Figure 1. The investigated problem: Is it possible to compute the 3D shape of a textured surface from such a single view using an uncalibrated camera?

This problem is clearly equivalent to multi-plane based camera pose estimation [16, 19]. According to Poncelet's theorem, given the image of a single texton in some Euclidean representation (i.e., under the assumption of square pixels), estimating the camera pose is ill-posed, as the locus of possible camera centres is a circle orthogonal to the image plane[1]. The intersection of the circle's supporting plane and the image plane is a straight line known as the *center line* such that for every point along it there exist a focal length and a camera pose which are consistent with this texton. It is well known that transformation from a planar scene to the image plane is a homography of the projective plane. It has been shown in [6] that the center line can be estimated from such a homography. As a result, using several textons will make the estimation of the camera parameters possible since different center lines may cross, and thus determine the location of the principal point (and consequently the focal length and the camera pose.) Accord-

[1] A movie which illustrates Poncelet's theorem is available at www.irit.fr/∼Pierre.Gurdjos/ECCV2002/

ing to this intuition, it seems that uncalibrated SFT is well-posed. However, this is not true when the textons are small, which is precisely a SFT assumption, because the estimation of a homography using a small texton is *ill-conditioned* [11].

As a remedy to this, we exploit the well known fact that the perspective projection model can be locally approximated by the scaled orthographic projection model. This provides good approximation to the imaging process with increasing accuracy for distant scene samples projected closer to the principal point. These assumptions are often satisfied in practice and permits the estimation of the associated affine transformation that is far more stable manner than the local homography (as exploited in other computer vision problems [10]).

The first contribution of our work is to provide a stable estimation of the depth and of the normal at each texton using an explicit piecewise planar affine model. A closed form solution is presented where, for both the calibrated and uncalibrated cases, orientation is given up to a two-fold ambiguity. Depth by contrast is recovered uniquely (where in the uncalibrated case it is known up to a scale factor).

Our second contribution extends the above process to joint 3D reconstruction and focal length estimation. The recovery of both the depth, which depends on the focal length, and of the normal (up to a two-fold ambiguity) at each patch is a redundancy which we exploit in order to estimate the focal length, and then by extension to recover the orientation and depth of the surface at each patch. Both these contributions are validated by experiments performed on synthetic and real images.

The structure of this paper is as follows. In Section 2 relevant works from the SFT literature are outlined. In Section 3 we discuss our method for computing the 3D shape of a textured surface from a single view given that the focal length of the camera is known. A generalization of this contribution to the case of an uncalibrated camera is done in Section 4, making an explicit estimation of the focal length in conjunction with shape. In Section 5 our work is summarised with reference to some possible future developments.

2. State of the Art

In the following section we briefly discuss papers from the SFT literature. To our knowledge however no up-to-date survey covering SFT is available at present.

Texture is known to be a strong cue for the perception of depth, and has been studied for a considerable time [5]. SFT is a classical 3D reconstruction process which essentially requires only a single image view. Two different approaches have arisen. When the texture is considered 'natural', statistical methods and descriptors are often used [18]. When the texture has been 'artificially' created, the fact that it is often repeating in nature suggests geometric approaches may be preferable [4]. The statistical methods rely often on a variety of densities estimators [1], whereas the geometric ones aim at estimating 2D transformations, which are homographies for perspective planar textures [8], but affinities have also been considered in some cases [14]. As with other 3D reconstruction techniques such as Shape-From-Shading or photometric stereo, SFT usually produces a normal field, and not 3D depth (which contrasts with our method). An additional step is then needed to recover shape, known as *normal integration* [9].

Recently a number of new SFT approaches have appeared. In [11], Lobay and Forsyth propose an automatic technique for textons detection and fronto-parallel appearance estimation. This work is fairly close to ours, except that they use orthographic cameras. In [12], Loh and Hartley's work does not require the texton pattern to be imaged in advance. This is very promising even if few results are shown. In [17], the combination of texture and shading is shown to be enough to avoid, in most cases, the ambiguity on the normal field estimation. We are interested in artificial textures, but unlike most of the previous works, we consider SFT as a multi-plane based camera pose estimation [16, 19].

3. Piecewise Planar Scaled Orthography

3.1. Camera Model

The imaging process we consider involves a *projective camera*, whose focal length f is the sole unknown (other intrinsics have canonical values *e.g.*, the principal point is at the origin). In the special case where the world frame coincides with the camera frame, its projection matrix writes:

$$\mathbf{P} = \mathrm{diag}(f, f, 1)\,[\mathbf{I}_{3\times 3} \mid \mathbf{0}_3]\,. \tag{1}$$

With regard to *small objects at a distance*, it is known that local affine approximations of $\mathbf{P}$ may provide more stable numerical solutions to the problem of computing the object pose.

In the immediate neighbourhood of any texton, the local affine approximation of the projective camera $\mathbf{P}$ acts like a *scaled orthographic camera* with projection matrix:

$$\mathbf{P}_j^{\mathrm{so}} = \begin{bmatrix} \alpha_j & 0 & 0 \\ 0 & \alpha_j & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{e}_1^\top \\ \mathbf{e}_2^\top \\ \mathbf{e}_4^\top \end{bmatrix}, \tag{2}$$

where $\alpha_j = f/d_j$ is the *scale factor*, j is the index of the particular texton, d_j is the depth at the patch's centroid $\mathbf{p}_j$ and $\mathbf{e}_k$ denotes the k^{th} column of $\mathbf{I}_{4\times 4}$. The model (2) is essentially a first-order approximation of (1), see [15, 7].

3.2. The Problem of Recovering the Pose of a Patch

The pose of patch j is given by $\mathbf{T}_j$, the 3D rigid transformation mapping a local world frame attached to its sup-

porting plane to the camera frame:

$$\mathbf{T}_j = \begin{bmatrix} \mathbf{R}_j & \mathbf{t}_j \\ \mathbf{0}_3^\top & 1 \end{bmatrix}$$

where $\mathbf{R}_j$ and $\mathbf{t}_j$ represent the rotation and translation components.

As we treat patches as *small objects at a distance*, the patch-to-texton 2D-transformation, denoted $\mathbf{A}_j$, mapping the patch j to its image can be well approximated by the scaled orthographic camera (2). Without loss of generality, take the supporting plane of patch j to be defined at $z = 0$ w.r.t. the local world frame and the patch's centroid at the origin. As a result the patch-to-texton transform $\mathbf{A}_j$ is affine and decomposed as

$$\mathbf{A}_j = \begin{bmatrix} \alpha_j \hat{\mathbf{R}}_j & \alpha_j \hat{\mathbf{t}}_j \\ \mathbf{0}_3^\top & 1 \end{bmatrix}$$

where $\hat{\mathbf{R}}_j$ denotes the top left 2×2 submatrix of $\mathbf{R}_j = [r_{kl}]_{(k,l)\in[1,3]^2}$, and $\hat{\mathbf{t}}_j$ denotes the top 2×1 elements of $\mathbf{t}_j$. Given $\mathbf{A}_j = [a_{kl}]_{(k,l)\in[1,3]^2}$, it can be shown that a *two-fold solution* can be directly obtained for both the pose $(\mathbf{R}_j, \mathbf{t}_j)$ and the scale factor α_j. We achieve this as follows. Introducing two unknown variables $\beta = \alpha_j r_{13}$ and $\gamma = \alpha_j r_{23}$, we can solve for β and γ as the solution to the following 2^{nd} order polynomial system:

$$\begin{cases} a_{11}^2 + a_{12}^2 + \beta^2 - \left(a_{21}^2 + a_{22}^2 + \gamma^2\right) = 0, \\ [a_{11}, a_{12}, \beta]\,[a_{21}, a_{22}, \gamma]^\top = 0. \end{cases}$$

This leads to two real solution pairs for β and γ $(\beta, \gamma) = \pm(b, c)$ for some real (b, c), and from this the patch pose is easily recovered:

$$\mathbf{R}_j = \frac{1}{\alpha_j} \begin{bmatrix} \mathbf{u}^\top \\ \mathbf{v}^\top \\ 1/\alpha_j (\mathbf{u} \times \mathbf{v})^\top \end{bmatrix}, \quad \mathbf{t}_j = \alpha_j^{-1}(a_{13}, a_{23}, f)^T \tag{3}$$

where $\mathbf{u} = [a_{11}, a_{12}, \beta]^\top$, $\mathbf{v} = [a_{21}, a_{22}, \gamma]^\top$, and $\alpha_j = \|\mathbf{u}\|^{-1}$. Notice that rotation component $\mathbf{R}_j$ and scale factor α_j are recoverable without knowing f, showing that planar orientation can be recovered with the scaled orthographic approximation *in the uncalibrated setting*. The two solutions for β and γ lead to a two-fold ambiguity in $\mathbf{R}_j$, but a single solution for α_j and $\mathbf{t}_j$. Now if $\boldsymbol{\pi}_j$ denotes the 4-vector of dual coordinates of the supporting plane of patch j, then it can be seen that

$$\boldsymbol{\pi}_j = \begin{bmatrix} \mathbf{n}_j^\top \\ -\mathbf{t}_j^\top \mathbf{n}_j \end{bmatrix} \tag{4}$$

where $\mathbf{n}_j = [\beta/\alpha_j, \gamma/\alpha_j, \det \hat{\mathbf{R}}_j]^\top$ denotes the normal in the camera's reference frame. Due to the ambiguity mentioned above, a two-fold solution clearly exists for $\boldsymbol{\pi}_j$.

3.3. Calibrated Orientation Disambiguation

Although we have a two-fold orientation ambiguity, the additional per-patch depths clearly provides us with redundant information to resolve this. Recall that most of the existing SFT techniques only compute the normals, which then requires a normal integration step. In our framework disambiguating the normals can be done far more easily. We make the assumption that neighbouring patch centroids $\mathbf{p}_k$ approximately sample the tangent plane at patch j (which is valid if the patch spatial separation is low). Thus, for a reconstructed patch the normal $\mathbf{n}_j^u$ is selected with maximal depth/normal agreement given by $S_k\left(\mathbf{n}_j^u\right) = \left|(\mathbf{p}_k - \mathbf{p}_j)^\top \mathbf{n}_j^u\right|$. In the presence of noise, the redundancy from multiple neighbours can be used to robustly disambiguate orientation (e.g. taking the normal with maximal votes these. One could further include 2^{nd} order agreement over the normal field, although compared with this local approach it has proven unnecessary. As explained, violations may occur, for example due to surface discontinuities. These are not treated in this paper, but should be handled by integrating additional discontinuity cues.

3.4. Reconstruction Experiments

In this section we show some simple reconstruction experiments conducted to validate or SFT method in the calibrated setting.

3.4.1 Synthetic Setup

Empirical synthetic studies to assess reconstruction quality in the presence of noise and differing scene conditions are now presented. A simple synthetic scene was constructed involving a cylinder quantised into regular squared grids of varying size to simulate the effect of varying texton densities. Three surfaces are shown in Fig. 2 a-c. We compute the images of these surfaces using a constant focal length (we use $f = 500$ with distances expressed in pixels), axis aligned with the camera's y axis, radius set to a constant $r = 500$. The surfaces are positioned at varying values of $\bar{d}$, denoting the *mean depth* of the scene; thus degree of perspective distortion is strong at short range, but disappears at long range.

The quality of the reconstructed patch orientations was evaluated with the RMS angular error (in degrees) in surface normals. Fig. 4 shows the trend of RMSE w.r.t. the number N of textons, for different values of $\bar{d}$. For a fixed value of $\bar{d}$, the error decreases for N, with confirms the assumption that the approximation improves with smaller patches. On the other hand, for a given number N of textons, RMSE decreases with $\bar{d}$, confirming the scaled orthographic approximation is better justified at longer distances.

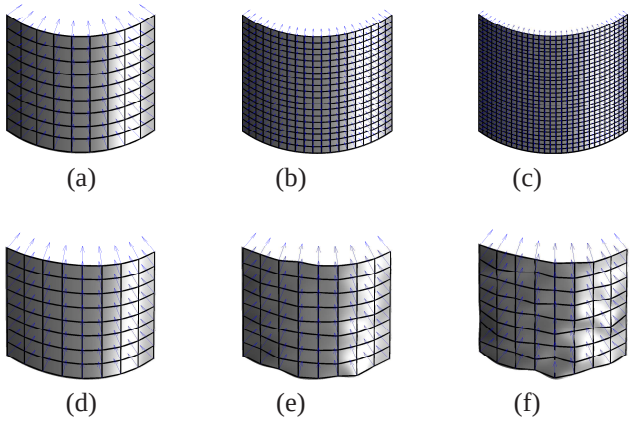


Figure 2. Test cylinders quantised into grids of (a) 10×10, (b) 20×20 and (c) 30×30 squared patterns. (d) Surface and normals recovered at short range ($\bar{d}/f = 2.5$) with no noise. (e,f) Reconstructions with measurement noises (additive Gaussian with standard deviations $\sigma = 0.1$ and $\sigma = 0.2$) respectively

To test the stability of reconstruction w.r.t. measurement errors, varying amounts of zero-mean Gaussian noise was added to each vertex's image location (from which the plane-to-view transforms are computed.) The three results shown in Fig. 2 correspond to zero noise, $\sigma = 0.02\ \%$ (Fig. 2-b) and $\sigma = 0.04\ \%$ (Fig. 2-c) of the image size, i.e., respectively $\sigma = 0.1$ and $\sigma = 0.2$ for a 512×512 image. In spite of noise, the surface's shape is recovered well. A detailed evaluation is presented in Fig. 3. Fig. 3-a and b denotes the depth (relative to the scene's depth range) and patch orientation error as a function of $\bar{d}$, performed at multiple noise levels. We can see here that there is trade-off between further distances (supporting the scaled orthography assumption) and shorter distances that reduce the influence of noise. We contrast this with recovering orientations based on texton homography decomposition [19] in Fig. 3-c. In the noise-free case, that method is clearly superior, however as noise is introduced and at further distances, we see our method is more successful.

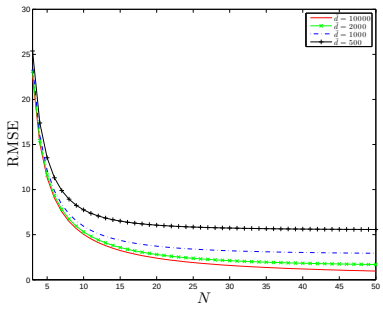


Figure 4. RMS angular error with respect to ground truth normals with varying texton density and mean depth.

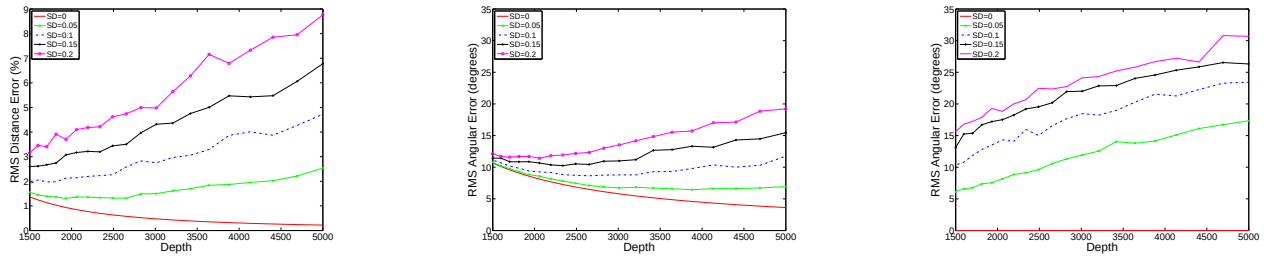
3.4.2 Real Setup

A second experiment we present was performed on a simple real scene depicting a surface with circular (non-regularly) printed textons (see Fig. 5-a). These are easily detected automatically (within a bounding region of interest manually specified) and the local transforms were refined using standard gradient-based image registration. The texton neighbourhood system was constructed with the edges of a simple Delaunay triangulation at the texton centres (as in all examples we present). More sophisticated strategies may of course be needed to develop better neighbourhood systems to handle, for example surface discontinuities and occlusions but this is left for now to future work.

The patch poses are estimated and shown in Fig. 5-c, displayed as orientated quad patches. The depths and normals appear very consistent, except for a clear outlier. A dense reconstruction was then performed using a robust interpolation strategy. For this we use the thin plate spline (TPS) with control points defined at the patch centres and a re-rendering showing the surface from another viewpoint is shown in Fig. 5-d. Ground truth depths were obtained with hand labeled stereo correspondences at the texton centres, resulting in an RMS depth error of 3.5% w.r.t. the depth of the enclosing volume.

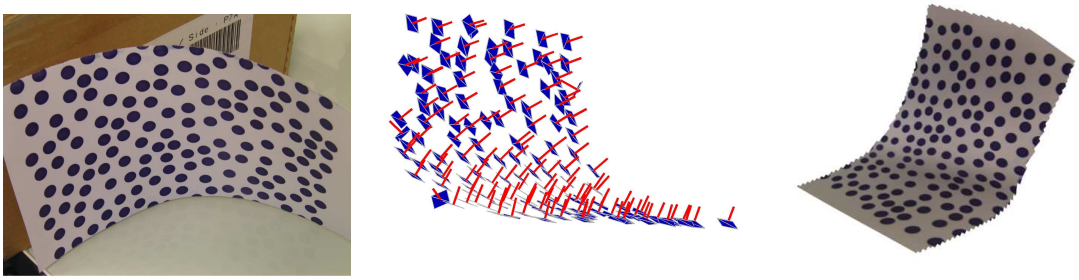
4. SFT with Uncalibrated Cameras

In this section we now generalise the methodology presented in Section 3 to the case where the camera is uncalibrated (with the focal length as the only unknown intrinsic). We propose two methods for calibrating f. The first is to estimate f using a criterion based on the normal integrability. After discussing its merits, we propose an alternative method which is considerably more reliable under near-affine viewing conditions. We also show how the nor-



(a) (b) (c)

Figure 3. Synthetic reconstruction results. (a) Depth and (b) orientation error as a function of mean scene depth. (c) Comparison to orientation reconstruction by homography decomposition



(a) Input image. (b) Recovered surface. (c) Resynthesized view.

Figure 5. (a) Image of a surface non-regularly printed with textons. (b) Recovered texton depths and normals (the patches are represented by oriented squares). (c) Re-rendered view of the reconstructed surface.

mal field can be disambiguated in the uncalibrated setting.

4.1. Focal Length by Integrability of Normal Field

Suppose first the normals have already been disambiguated. Integrating a normal field is a classical computer vision problem. The equation of normal integration expresses that the variation in depth along any closed loop in the image is equal to zero (the depth gradient field must be a zero-curl field), a property which has been used for many other purposes than for normal integration [12]. Let us first use as a measure of consistency between the depth field $d(x,y)$ and the normal field $\mathbf{n}(x,y) = [n_x(x,y), n_y(x,y), n_z(x,y)]^\top$ the following equality, which holds under perspective projection [2]:

$$\boldsymbol{\nabla}\delta = [r,s]^\top, \tag{5}$$

where $\delta = \ln|d|$ and:

$$\begin{cases} r = -\dfrac{n_x}{x\,n_x + y\,n_y + f\,n_z}, \\ s = -\dfrac{n_y}{x\,n_x + y\,n_y + f\,n_z}. \end{cases}$$

The definitions of r and s explicitly depend on f. Therefore, it seems that the following criterion is sensitive to f:

$$\mathcal{C}_1(f) = \sum_j \left\| \boldsymbol{\nabla}\delta(\mathbf{q}_j) - [r_j, s_j]^\top \right\|^2, \tag{6}$$

where $\mathbf{q}_j$ designates the centroid of texton j, and the sum is carried out on all the textons. Anything is known in (6), except f and the values $\boldsymbol{\nabla}\delta(\mathbf{q}_j)$, which can be estimated using Taylor expansions: $\delta(\mathbf{q}_k) = \delta(\mathbf{q}_j) + (\mathbf{q}_k - \mathbf{q}_j)^\top \boldsymbol{\nabla}\delta(\mathbf{q}_j)$.

We tested the criterion $\mathcal{C}_1$ on images with very low perspective distortion, and surprisingly, it seems to fail. $\mathcal{C}_1$ is quasi-insensitive to f and, therefore, the estimation of f is ill-conditioned. Using, once again, the scaled orthographic model, Equation (5) reduces to the following [9, 2]:

$$\boldsymbol{\nabla} d = -\frac{1}{\alpha}\left[\frac{n_x}{n_z}, \frac{n_y}{n_z}\right]^\top. \tag{7}$$

An approximate version of $\mathcal{C}_1$ is then:

$$\mathcal{C}_2(f) = \sum_j \left\| \boldsymbol{\nabla} d(\mathbf{q}_j) + \frac{1}{\alpha_j}\left[\frac{n_{x,j}}{n_{z,j}}, \frac{n_{y,j}}{n_{z,j}}\right]^\top \right\|^2. \tag{8}$$

Anything is known in (8), except f and the values $\boldsymbol{\nabla} d(\mathbf{q}_j)$,

which can be estimated from the following linear systems:

$$d_k = d_j + (\mathbf{q}_k - \mathbf{q}_j)^\top \boldsymbol{\nabla} d(\mathbf{q}_j), \tag{9a}$$

$$d_l = d_j + (\mathbf{q}_l - \mathbf{q}_j)^\top \boldsymbol{\nabla} d(\mathbf{q}_j), \tag{9b}$$

where k and l are indices for two neighbouring textons of texton j. Since the depth $d = f/\alpha$ is proportional to f, $\mathcal{C}_2$ can be rewritten:

$$\mathcal{C}_2(f) = \sum_j \left\| f\,[\lambda_j, \mu_j]^\top + \frac{1}{\alpha_j}\left[\frac{n_{x,j}}{n_{z,j}}, \frac{n_{y,j}}{n_{z,j}}\right]^\top \right\|^2. \tag{10}$$

where the λ_j and μ_j are obtained from (9a)-(9b), which is a Cramer system as soon as $\mathbf{q}_j = [x_j, y_j]^\top$, $\mathbf{q}_k = [x_k, y_k]^\top$ and $\mathbf{q}_l = [x_l, y_l]^\top$ are non-colinear:

$$\lambda_j = \frac{(y_l - y_j)(1/\alpha_k - 1/\alpha_j) - (y_k - y_j)(1/\alpha_l - 1/\alpha_j)}{(x_k - x_j)(y_l - y_j) - (y_k - y_j)(x_l - x_j)}, \tag{11a}$$

$$\mu_j = \frac{(x_k - x_j)(1/\alpha_l - 1/\alpha_j) - (x_l - x_j)(1/\alpha_k - 1/\alpha_j)}{(x_k - x_j)(y_l - y_j) - (y_k - y_j)(x_l - x_j)}. \tag{11b}$$

From (10), we can easily express the optimal value f^* of f:

$$f^* = -\frac{1}{\sum_j \lambda_j^2 + \mu_j^2} \sum_j \frac{\lambda_j n_{x,j} + \mu_j n_{y,j}}{\alpha_j n_{z,j}}. \tag{12}$$

We see in (11a) and (11b) that the coefficients λ_j and μ_j tend towards zero when the global perspective effect vanishes, since all the scale factors α_j tend to the same value. Obviously, this now makes the estimation (12) of f^* ill-conditioned.

4.2. Uncalibrated Orientation Disambiguation

We now show how the normals can be disambiguated. For a given patch j, we know from Section 3.2 that there are two normal solutions $\mathbf{n}_j^1 = 1/\alpha_j[b, c, h]^\top$ and $\mathbf{n}_j^2 = 1/\alpha_j[-b, -c, h]^\top$ for some known b, c and h. Using the scaled orthographic model, Equation (7) then reads:

$$\boldsymbol{\nabla} d(\mathbf{q}_j) = \pm\frac{1}{\alpha_j}\left[\frac{b}{h}, \frac{c}{h}\right]^\top. \tag{13}$$

Recall that from the linear system (9a)-(9b), we can deduce $\boldsymbol{\nabla} d(\mathbf{q}_j) = f\,[\lambda_j, \mu_j]^\top$ as soon as texton centroids $\mathbf{q}_j$, $\mathbf{q}_k$ and $\mathbf{q}_l$ are non-colinear. Equation (13) can thus be rewritten:

$$f\,\boldsymbol{v} = \pm\boldsymbol{w}, \tag{14}$$

where the two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ are known and do not depend on f, which nevertheless is involved in Equation (14). Since the estimation of f supposes that the disambiguation of the normals is already carried out, as seen in Section 4.1, the problem of disambiguating $\mathbf{n}_j$, which is that of finding the right sign in (14), looks like a vicious circle. But, knowing that $f > 0$, it is reasonable to compute $s = \text{sign}\,(\boldsymbol{v}^\top \boldsymbol{w})$ and to conclude:

$$\mathbf{n}_j = \frac{1}{\alpha_j}[-sb, -sc, h]^\top.$$

4.3. Focal Length by Maximising Continuity

The second approach we propose for calibrating f is to find the focal length which maximises the continuity of the depth field. Consider two adjacent patches j and k in 3D space. To recover the focal length, first assume we have at our disposal the homogeneous pixel coordinates $\mathbf{q} \in \mathbb{R}^3$ of a point lying on the image of the intersection of the two corresponding supporting planes. The 3D line $\mathcal{L}_\mathbf{q}$ back-projected from $\mathbf{q}$ under the projective camera $\mathbf{P}$ in (1) may be written as a dual Plücker 4×4-matrix

$$\mathbf{L}^* = \mathbf{P}^\top [\mathbf{q}]_\times \mathbf{P},$$

where $[\mathbf{q}]_\times$ denotes the usual skew-symmetric matrix of $\mathbf{q}$.

The 3D point $\mathbf{X}_j$ at which $\mathcal{L}_\mathbf{q}$ meets the supporting plane of patch j is

$$\mathbf{X}_j = \mathbf{L}\boldsymbol{\pi}_j \tag{15}$$

where $\mathbf{L}$ is the primal Plücker 4×4-matrix (i.e. *dual* to $\mathbf{L}^*$) and $\boldsymbol{\pi}_j$ the 4-vector (4) of dual coordinates of the supporting plane of patch j. A similar equation to (15) holds for the supporting plane of patch k, which meets $\mathcal{L}_\mathbf{q}$ in $\mathbf{X}_k$. The question now remains of how to recover f. As $\mathcal{L}_\mathbf{q}$ intersects $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$ at the same 3D point, any Cartesian (*i.e.*, normalised) coordinate of $\mathbf{X}_j$ and $\mathbf{X}_k$ must be equal. Hence, using this constraint, we can easily derive a degree-2 polynomial equation in f denoted by $c_2 f^2 + c_1 f + c_0 = 0$, whose coefficients have the form

$$\begin{cases} c_2 = (\alpha_k - \alpha_j)\; g_2, \\ c_1 = \alpha_j \alpha_k\; g_{1jk} + \alpha_j g_{1j} + \alpha_k g_{1k}, \\ c_0 = \alpha_j \alpha_k\; g_0, \end{cases}$$

where $g_* = g_*(\mathbf{n}_j, \mathbf{n}_k, \hat{\mathbf{t}}_j, \hat{\mathbf{t}}_k)$ denote functions not depending on α_j or α_k. If we assume that $d_j \approx d_k$, then $\alpha_j \approx \alpha_k$ and, even this is not strictly true, we can understand, looking at c_1 and c_2, why in practice $|c_2| \ll |c_1|$. Hence, a first root of the quadratic equation is $f_1 \approx -c_0/c_1$. Knowing that $f_1\, f_2 = c_0/c_2$, we can conclude that the other solution is $f_2 \approx -c_1/c_2$. Since $f_1 + f_2 = -c_1/c_2$, we notice that $|f_1| \ll |f_2|$. Thus we can recover a single solution to be the smallest of the two roots. We have arrived at a minimal solution for f requiring two adjacent textons. By exploiting all texture information we can now proceed to find the single optimal focal length. Currently we adopt a simple robust strategy: we first reject the focal lengths from

any texton pairs whose normals are at an angle greater than 20 degrees. These occurrences are either caused by surface discontinuities or erroneous normals from either patch, and should be discarded. Then we robustly compute the optimal focal length as the median of all remaining focal lengths.

4.4. Experiments

We have evaluated our second focal length estimation method using some real world examples and these are now presented. Fig. 6-a illustrates one case. The image is 1200 $\times$ 1600 pixels with a focal length of 1274 pixels comprising an 8 $\times$ 7 lattice of square textons. The patch-to-texton mappings were achieved by hand labeling. Superimposed are the recovered normals (red) and ground truths (green) obtained from planes fitted to triangulated stereo correspondences. The accuracy here is clear; the RMS angular error is 2.3 degrees. To assess the effect of measurement noise on both focal length estimation and reconstruction, the corner positions were perturbed with Gaussian noise and Fig. 6-b plots the relative error in focal length against correspondence noise. In the noise-free case we have a relative error of 9.1% and as one can see, the focal estimate accuracy degrades gracefully with increased noise, suggesting the method is stable in real conditions. We also notice that with increased noise the recovered focal length (and consequently the mean depth of the surface) becomes underestimated. This makes sense; since with increased noise the normal estimates tend towards random, and the continuity criterion becomes best satisfied when samples lie at the plane $z = 0$. The corresponding errors in orientation and depth are presented in Fig. 6-c.

A second example is presented in Fig. 7-a where we have an image of a textured deformed shirt. The textons were marked by hand, the focal length is 1425 pixels and ground truth data was acquired with manual stereo triangulation. A smooth TPS surface interpolates these points to serve as ground truth depth and orientation data. Fig. 7-b shows the orientation and depth reconstruction accuracy. The range of errors here is fairly broad, chiefly due to the local nonrigidity of the cloth and noise from manual labeling. Even so, we still obtain a focal length of 1749 (a 22% relative error). Shape appears to have been recovered well. Fig. 7-c shows a smooth surface reconstructed from the texton centres with lines indicating the texton normals. Fig. 7-d shows the ground truth surface with texton centres marked.

5. Conclusion and Perspectives

In this paper we have revisited the SFT problem using a single view to reconstruct surfaces using instances of projected textons. Our first contribution is to show that when the camera's intrinsics are known, the 3D shape of the surface can be computed *without normal integration*, by approximating global perspective by local scaled orthography. This is suitable when the textons are small where there is insufficient local perspective to resolve the per-texton homographies. Our second contribution is to generalize this result to the uncalibrated setting using the redundancy between depths and normals. Our method makes this possible because only the depths, rather than orientation depend on the focal length.

One strong possible application of our work is to undistort the camera images of curled documents [13]. Since the automatic detection of characters in an image may provide a rather dense estimation of the 3D shape, it will be possible to flatten the document (for restoration or improving readability) without needing strong geometric assumptions. Furthermore, with the ubiquity of text in many images, the ability to calibrate a camera from groupings of characters within a single image is an appealing prospect. In future we aim to handle surfaces with discontinuities and include automatic texton registration. Another possibility is to handle the two-fold ambiguity on the normal using programming heuristics such as those described in [3], rather than using the redundancy between depths and normals. Finally, we aim to generalize the theory of our work to the more difficult, but more realistic and applicable situation when the frontoparallel pattern appearances are unknown s[12].

References

[1] R. Bajcsy and L. Lieberman. Texture gradient as a depth cue. *Computer Graphics and Image Processing*, 5(1):52–67, 1976. 2

[2] J.-D. Durou and F. Courteille. Integration of a Normal Field without Boundary Condition. In *Proceedings of PACV (workshop of ICCV 2007)*, 2007. 5

[3] A. Ecker, A. D. Jepson, and K. N. Kutulakos. Semidefinite Programming Heuristics for Surface Reconstruction Ambiguities. In *Proceedings of ECCV 2008 (volume 1)*, pages 127–140, 2008. 7

[4] J. Garding. Shape from texture for smooth curved surfaces in perspective projection. *Journal of Mathematical Imaging and Vision*, 2(4):327–350, 1992. 2

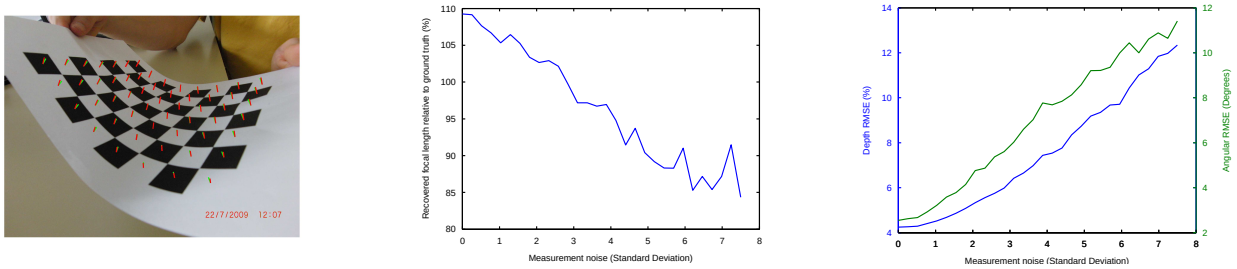
[5] J. J. Gibson. *The Perception of the Visual World*. Houghton Mifflin, 1950. 2

[6] P. Gurdjos, A. Crouzil, and R. Payrissat. Another Way of Looking at Plane-Based Calibration: the Centre Circle Constraint. In *Proceedings of ECCV 2002 (volume 4)*, pages 252–266, 2002. 1

[7] R. Horaud, F. Dornaika, B. Lamiroy, and S. Christy. Object Pose: The Link between Weak Perspective, Paraperspective, and Full Perspective. *International Journal of Computer Vision*, 22(2):173–189, 1997. 2

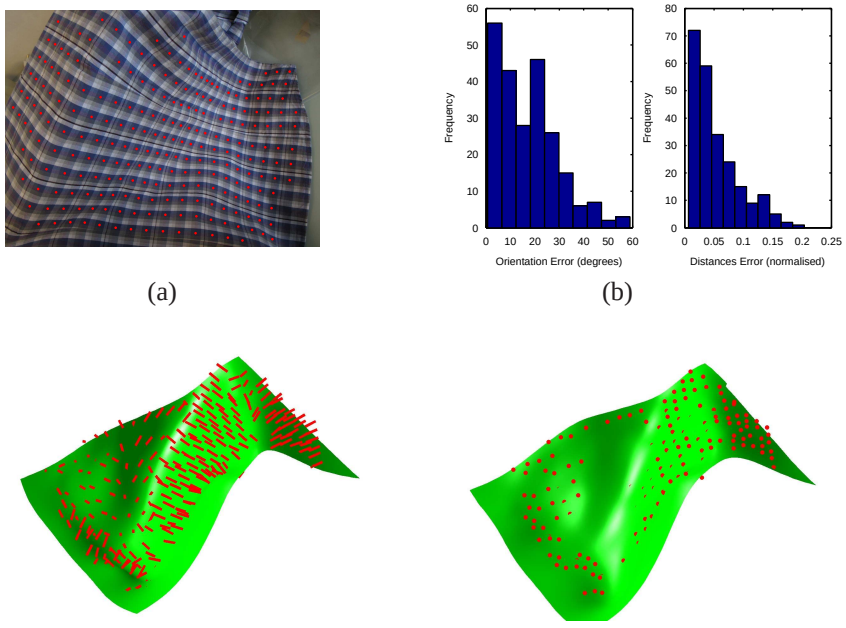
[8] K. Ikeuchi. Shape from Regular Patterns. *Artificial Intelligence*, 22(1):49–75, 1984. 2

[9] P. Kovesi. Shapelets Correlated with Surface Normals Produce Surfaces. In *Proceedings of ICCV'05*, pages 994–1001, 2005. 2, 5



(a) (b) (c)

Figure 6. Results of focal length estimation. (a) An example of textured surface. (b) Effect of correspondence noise on focal length estimation and (c) uncalibrated reconstruction.



(a) (b)

(c) (d)

Figure 7. Uncalibrated reconstruction of textured cloth. (a) Single view with texton centres marked by dots. (b) Reconstruction performance in terms of orientation (Left) and distance (Right). (c) Reconstructed surface with texton normals shown. (d) Ground truth surface from stereo with textons marked by dots.

[10] D. Lingrand. Particular Forms of Homography Matrices. In *Proceedings of BMVC'00 (volume 2)*, pages 596–605, 2000. 2

[11] A. Lobay and D. A. Forsyth. Shape from texture without boundaries. *International Journal of Computer Vision*, 67(1):71–91, 2006. 2

[12] A. M. Loh and R. Hartley. Shape from non-homogeneous, non-stationary, anisotropic, perspective texture. In *Proceedings of BMVC'05*, pages 69–78, 2005. 2, 5, 7

[13] S. Lu, B. M. Chen, and C. C. Ko. A partition approach for the restoration of camera images of planar and curled documents. *Image and Vision Computing*, 24(8):837–848, 2006. 7

[14] Y.-I. Ohta, K. Maenobu, and T. Sakai. Obtaining surface orientation from texels under perspective projection. In *Proceedings of IJCAI'81*, pages 746–751, 1981. 2

[15] S.-C. Pei and L.-G. Liou. Finding the motion, position and orientation of a planar patch in 3D space from scaled-orthographic projection. *Pattern Recognition*, 27(1):9–25, 1994. 2

[16] P. Sturm. Algorithms for Plane-Based Pose Estimation. In *Proceedings of CVPR'00*, pages 1010–1017, 2000. 1, 2

[17] R. White and D. A. Forsyth. Combining Cues: Shape from Shading and Texture. In *Proceedings of CVPR'06*, 2006. 2

[18] A. P. Witkin. Recovering Surface Shape and Orientation from Texture. *Artificial Intelligence*, 17(1–3):17–45, 1981. 2

[19] Z. Zhang. Camera Calibration By Viewing a Plane From Unknown Orientations. In *Proceedings of ICCV'99*, pages 666–673, 1999. 1, 2, 4