Vision, Modeling, and Visualization (2010)

Locally Planar and Affine Deformable Surface Reconstruction from Video

T. Collins[1,2] and A. Bartoli[1]

[1]Universitï£¡ d'Auvergne, Clermont-Ferrand, France [2]University of Edinburgh, UK
Toby.Collins@gmail.com Adrien.Bartoli@gmail.com

Abstract

We present new solutions to tackle the problem of reconstructing a deforming surface viewed in monocular videos without a template, by exploiting the fact that many deforming surfaces are on the local scale approximately rigid and planar. The reconstruction task can then be seen, from bottom up as first multi-plane based pose estimation then dense surface reconstruction from planar samples. In practice there are major obstacles to overcome. In this paper we specifically target computing stable orientation estimates at small surface regions from interview image motion. We achieve this using local affine projection models which are stable and accurate when local perspective effects are small. Our core theoretical contributions are closed form solutions to multiview orthographic planar pose estimation in both the minimal and overdetermined cases. We use this to efficiently construct a weak deformable template; an undirected graph with nodes holding the surface's local planar structure and edges denoting physical deformation constraints. The template can then be used to recover dense 3D shape very efficiently from affine image motion via unambiguous planar pose estimation combined with surface regularisation.

1. Introduction

Recovering the 3D shape and motion of nonrigidly moving bodies from monocular image data remains one of the most sought after goals in computer vision. The Non-Rigid Structure From Motion (NR-SFM) paradigm uses motion detected on the camera's image plane to recover 3D information. Currently there are two broad NR-SFM categories. Category 1: Template-Based, requires a model of the surface geometry (*i.e.* a 3D template.) This is assumed to be known prior to reconstruction. Category 2: Templateless, is the more recent and attempts to recover geometry and motion with no such template. This is a considerably more challenging problem. Even if a geometric template is known the problem is intractable without additional assumptions. Methods in both categories can be separated by what particular assumptions are made. Currently the two most common are (a) statistical low rank assumptions and (b) physical assumptions. This work fits into category 2(b). We exploit constraints arising from the assumption of approximate local rigidity; a characteristic found for objects made from a broad range of materials such as paper, cloth and plastics. A few recent works have pursued this direction. In [VSTF09] homographies from planar perspective projection were used to recover surface normals, which was followed by enforcing surface continuity to recover 3D shape. However, the core physical assumption is that the deformable surface is planar on the *local* scale. It is well known that homography estimation from small image regions is ill-conditioned [LF06], and perhaps should not be relied on for general deformable surface reconstruction. Instead in this paper we reject the idea of estimating local perspective transforms, and use the fact that the perspective model can be locally approximated by Scaled Orthographic (SO) models. This leads to affine interview transforms, which are estimated more stably than local homographies, yet retains good pose estimates in practice. This idea relates to recent work in Shape-from-Texture [CDGB10]. This can be considered a special case of plane-based SfM, but when the fronto-parallel planar appearance is known (*i.e.* it is *locally* template based.)

There are other template-based methods that use the closely related inextensibility constraint with considerable success [SF09, SSL10, PHB10, BHB*10, FXC09]. The recent convex problem formulation has marked a major step forward, however in the templateless case the problem is no longer convex. Inextensibility in conjunction with PCA-like shape models have also been considered [VSTF09, SUF08].

DOI: 10.2312/PE/VMV/VMV10/339-346

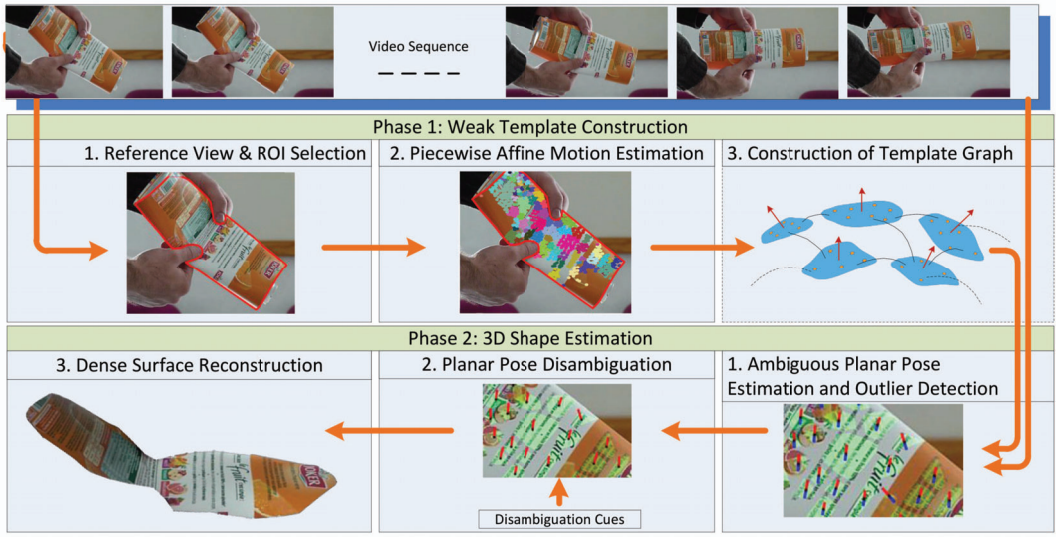


Figure 1: *Proposed NR-SFM Pipeline*

Very recently there has been some work in templateless inextensible surface reconstruction. In [WF06] fronto-parallel views of a surface's texture were recovered from orthographic views, and the resulting ambiguities then largely eliminated with the shading cue and global perspective. In complex illumination settings shading may be difficult to work with however. In this work we seek disambiguation using geometric and temporal constraints alone. Furthermore, frontoparallel planar views were computed based on [LF06], which uses exhaustive search. In [FXC09] inextensible surfaces are reconstructed from point correspondences using a novel factorisation-based approach. This appears promising, but required dense correspondences (such that the euclidean approximation to geodesic distances is reasonable), no mismatches, and the surface to be globally developable. At the time of submission we have become aware of concurrent work by Taylor *et al.* [TJK10]. Their's is similar in treating the reconstruction first as orthographic planar pose estimation. While their method of projected-length equations provides a closed form solution to planar structure from point samples, they then estimate planar pose with a second stage using nonlinear iterative least squares. For surfaces with many planes this may be time consuming, and may find only local minima corresponding to spurious solutions. Furthermore, their method requires 4 or more views for pose estimation. By contrast we present a method for closed form *structure and pose*. This covers the minimal case of 3 views. We believe this is the first method in existence to achieve this. Technically, [TJK10] differs by reconstructions based on a contiguous triangulation of feature points, and may suffer from noise and drop-off. Ours is based on clusters of freeform planar regions, and consequently may offer better stability and robustness.

The overarching pipeline we take from image sequences to reconstructed deforming 3D surfaces is illustrated in Fig. 1, and is broken into two broad phases. The first is template construction: modelling the surface's topology, local planar structure and appearance from interview image motion. The second is 3D shape estimation: exploiting the template to reconstruct shape in each view. Fully automatic template construction is certainly the harder process. The focus of this paper is not on topology estimation. Here we assume the surface is of disc topology and an unoccluded view is present in at least one reference frame. This is manually selected by a user with a corresponding Region of Interest (ROI).

The template building process is as follows. In stage 1 the reference frame and ROI is selected by the user. In stage 2, piecewise affine motion is estimated within the ROI over the image sequence. We derive this from point tracks, and assign these to spatially localised clusters. Each cluster collectively move according to the same affine motion. This is automatic, and provides (i) clusters robust to outlier tracks and (ii) it reveals the extent of the surface's local planarity. The problem is posed as a MRF-based segmentation, however we do not consider this a key contribution and defer exact details to the supplementary material. In stage 3, the deformable template is constructed from the clusters. The template is an undirected graph $T = (V, E)$, where each planar cluster defines a node $v_i \in V$. The template is used primarily to counter the problem that planar pose estimation from affine motion is inherently ambiguous: we have a 2-fold ambiguity per-plane due to Necker reversal. We associate with each node a binary Necker state, and the graph's edges E correspond to physical constraints acting between the nodes which serve to resolve the ambiguities. We call this a weak template, since it does

not correspond to a complete 3D surface template, but rather a locally-planar abstraction.

The planar structure of each node is estimated from its affine motion and an inversion of the planar projection process. In §3.1 we provide the theory for SO projection models which extends the theory given in [CDGB10] to the multi-view templateless setting. Affine motion of a plane under SO-projection holds its euclidean structure up to a 3-parameter 2D affine group, and recovering pose in all views becomes a problem of finding the euclidean upgrade. In §3.2 we provide our closed-form solutions to the minimal and overdetermined cases for single scale orthographic projection. We call this Orthographic Affine Decomposition (OAD). Our solutions are absent in the literature and has applicability beyond NR-SFM. In §3.3 we provide empirical results supporting OAD.

In §4 we present our 3D shape estimation process using the weak template. This is also divided into 3 stages. In Stage 1, ambiguous planar poses are estimated for each template node, using a closed form solution. Outliers can be detected based on a local consensus using neighbouring nodes. In Stage 2 (§4.1–4.2), planar poses are disambiguated using the weak template's edge constraints, derived from a model of local surface bending. In non-degenerate surface configurations this can only reduce the template's ambiguity to a global 2-fold. With also the assumption of temporal smoothness combined with one unambiguous frame, we can arrive at a unique solution across the video (§4.4). In stage 3 (§4.5), the unambiguous normals are used to recover a dense reconstruction, posed as a regularised system using a sparse normal field. In §4.6 we present results of our method using real image sequences, which is followed in §5 with concluding remarks, current limitations and future work.

2. Background: Template-Based Planar Pose Recovery with SO Cameras

We now review the theory of Scaled Orthographic planar pose estimation given affine motion between two projected views. Here the plane's orientation in the first view is known and normal aligned along the camera's z axis. We call this template-based pose recovery, since the euclidean structure of the plane in the first view is known. Suppose we are imaging a 3D surface $\mathcal{S}$ with a perspective camera with central projection matrix given by $\mathbf{P} = diag\,(f,f,1)\,[\mathbf{I}_{3\times3}|\mathbf{O}_3]$. We treat as unknown the focal length f and other projection parameters (principle point and skew) assumed known and their effects undone. While the model is globally perspective, the projection of small image regions can be well approximated by local affine models. A first order approximation to perspective projection $\psi(\mathbf{Pq})$, where $\mathbf{q} = [x,y,z,1]^\top$ is a point in homogeneous 3D coordinates in the camera's frame and $\psi\left((x_1,x_2,x_3)^\top\right) = (x_1/x_3, x_2/x_3, 1)^\top$, is given by the SO model: $\psi(\mathbf{Pq}) \approx \left[\text{diag}\,(\alpha,\alpha,0)\,\middle|\,[0,0,1]^\top\right]\mathbf{q}$. $\alpha_i = f/z$ denotes a local isotropic scaling factor. For planar projection, denote the transformation $\mathbf{T}_t = \begin{bmatrix} \mathbf{R}_t & \mathbf{t}_t \\ \mathbf{0}^\top & 1 \end{bmatrix}$ mapping a planar region defined at $z=0$ and centred at the origin into the camera's coordinate system at some time t. Let us also define the notation $\hat{\mathbf{Y}} = [\mathbf{Y}]_{2\times2}$ to mean taking the top left 2×2 submatrix of some matrix $\mathbf{Y}\in\mathbb{R}^{3\times3}$, and $\hat{\mathbf{v}} = [\mathbf{v}]_{2\times1}$ taking the top 2×1 elements of some vector $\mathbf{v}\in\mathbb{R}^{3\times1}$. The plane-to-image projection $\mathbf{A}_t$ is given by

$$\mathbf{A}_t = \begin{bmatrix} \alpha_t\hat{\mathbf{R}}_t & \alpha_t\hat{\mathbf{t}}_t \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{1}$$

It was shown in [CDGB10] that planar pose can be reconstructed by factoring $\mathbf{A}_t$ to give a solution unique in α_t and a 2-fold ambiguity in $\mathbf{R}_t$. This is of the form:

$$\mathbf{R}_t = \alpha_t^{-1}\begin{bmatrix} \hat{\mathbf{A}}_t & \gamma\begin{bmatrix} g \\ h \end{bmatrix} \\ \gamma\begin{bmatrix} k & l \end{bmatrix} & n_z \end{bmatrix} \quad \gamma\in\{-1,1\} \tag{2}$$

with $n_z = \alpha_t^{-1}\det\left(\hat{\mathbf{A}}_t\right)$. The solution is ambiguous up to a reflection about the z axis (*i.e.* a Necker reversal) denoted by the binary variable γ which we call the plane's *Necker state*. When γ is known let us define the unambiguous recovery of $\mathbf{R}_j$ using Eq.(2) with the notation $\mathbf{R}_j = \left[\hat{\mathbf{A}}_j,\gamma\right]_{3\times3}$: $\mathbb{R}^4\times\{-1,1\}\rightarrow\mathbb{S}^3$. Importantly the decomposition does not depend on the focal length, and so is applicable for uncalibrated perspective and orthographic cameras.

3. Template-Free Planar Pose from SO Views

We now generalise the theory of §2 to the multi-view templateless setting and present our closed-form solutions for single-scale orthographic planar pose estimation in the minimal and general $n\geq3$-view cases.

3.1. Multiview Affine Structure

The transform $\mathbf{A}_{ji}$ between two projected views i and j of a rigidly moving plane $\mathcal{P}$ under SO projection is given by:

$$\mathbf{A}_{ji} = \mathbf{A}_i\mathbf{A}_j^{-1} = \begin{bmatrix} \alpha_i\hat{\mathbf{R}}_i & \alpha_i\mathbf{t}_i \\ 0^T & 1 \end{bmatrix}\begin{bmatrix} \alpha_j\hat{\mathbf{R}}_j & \alpha_j\mathbf{t}_j \\ 0^T & 1 \end{bmatrix}^{-1} \tag{3}$$

Suppose we have n views of $\mathcal{P}$. Ignoring the translation terms, a $2n\times2n$ inter-view measurement matrix $\mathbf{M}$ can be constructed which factorises according to:

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}_2 & \hat{\mathbf{A}}_{21} & \cdots & \hat{\mathbf{A}}_{n1} \\ \hat{\mathbf{A}}_{12} & \mathbf{I}_2 & \cdots & \hat{\mathbf{A}}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{A}}_{1n} & \hat{\mathbf{A}}_{2n} & \cdots & \mathbf{I}_2 \end{bmatrix} = \begin{bmatrix} \alpha_1\hat{\mathbf{R}}_1 \\ \alpha_2\hat{\mathbf{R}}_2 \\ \vdots \\ \alpha_n\hat{\mathbf{R}}_n \end{bmatrix}\begin{bmatrix} \alpha_1^{-1}\hat{\mathbf{R}}_1^{-\top} \\ \alpha_2^{-1}\hat{\mathbf{R}}_2^{-\top} \\ \vdots \\ \alpha_n^{-1}\hat{\mathbf{R}}_n^{-\top} \end{bmatrix}^\top + \varepsilon \tag{4}$$

with ε denoting measurement noise. Consider the left $2n\times2$ factor $\begin{bmatrix} \hat{\mathbf{A}}_1^\top & \hat{\mathbf{A}}_2^\top & \cdots & \hat{\mathbf{A}}_n^\top \end{bmatrix}^\top = \mathbf{P}$, with $\hat{\mathbf{A}}_t = \alpha_t\hat{\mathbf{R}}_t$. Referring to Eq.(2), $\hat{\mathbf{A}}_t$ contains the 3D orientation of $\mathcal{P}$ at view t up to a 2-fold ambiguity. In the templateless case

the factor $\mathbf{P}$ is unknown. Suppose however we have an estimate of the left column-block of $\mathbf{M}$ (acquired for example by tracking from the first view: $\left[\begin{array}{cccc}\mathbf{I}_2 & \hat{\mathbf{A}}_{12}^\top & \cdots & \hat{\mathbf{A}}_{1n}^\top\end{array}\right] = \mathbf{B}$. Consider now as unknown the 2×2 plane-to-view projection term for view 1: $\hat{\mathbf{A}}_1 = \left[\begin{array}{cc} x_a & x_b \\ x_c & x_d\end{array}\right] = \mathbf{X}$. Given $\mathbf{X}$, $\mathbf{P}$ is now recoverable from $\mathbf{B}$ by $\mathbf{P} = \mathbf{B}\mathbf{X}$. $\mathbf{X}$ can therefore be seen as an *upgrading* matrix taking the multiview affine image structure $\mathbf{B}$ to 3D euclidean geometry held in $\mathbf{P}$. The rotation component of $\mathbf{X}$ denotes an arbitrary rotation on the support plane of $\mathcal{P}$, and so is uninformative for pose estimation at each view. This can be eliminated by for example clamping $x_b = 0$. Thus the matrix $\mathbf{B}$ contains the multiview pose structure up to a 3-parameter family (2D shear and anisotropic scale) and general 2^n-fold orientation ambiguity. Note that the affine structure was defined in a tracking setting with with respect to the first view, however this is not a requirement. Suppose we have access to the matrix $\mathbf{M}$ (with possibly missing entries.) $\mathbf{B}$ can be computed, up to the affine ambiguity, by taking the closest rank-2 decomposition of $\mathbf{M}$ using for example the SVD. This has the benefit of exploiting all measurement redundancy in $\mathbf{M}$.

Under SO approximation, the problem of euclidean pose estimation is that of finding $\mathbf{X}$ such that the matrix $\mathbf{B}\mathbf{X}$ decomposes with $\mathbf{B}\mathbf{X} = \left[\begin{array}{cccc}\alpha_1\hat{\mathbf{R}}_1^\top & \alpha_2\hat{\mathbf{R}}_2^\top & \cdots & \alpha_n\hat{\mathbf{R}}_n^\top\end{array}\right]^\top$. This is under-constrained given a single plane's affine structure; each $\alpha_t\hat{\mathbf{R}}_t$ has 4 DOF, thus for an arbitrary $\mathbf{X}$ we can find such a decomposition.

3.2. Upgrading with Orthographic Affine Decomposition (OAD)

In the orthographic case we assume $\alpha_t = \gamma\,\forall t$. This is a valid approximation to perspective projection when the relative change in depth of the plane is small when compared to the depth to the camera's centre. The world-to-image scaling is irrecoverable, yet does not affect orientation recovery and may be arbitrarily set to $\gamma = 1$. With no measurement noise, $\mathbf{X}$ and each 2×2 sub-block $\mathbf{B}_t$ of $\mathbf{B}$ are then related by:

$$\left[\begin{array}{cc}\mathbf{B}_t\mathbf{X} & \mathbf{a}_t \\ \mathbf{b}_t^T & c_t\end{array}\right] = \mathbf{R}_t \in \mathbb{S}^3$$

for some $\mathbf{a}_t, \mathbf{b}_t \in \mathbb{R}^{2\times 1}$ and scalar c_t. This implies $\mathbf{B}_t\mathbf{X}\mathbf{X}^\top\mathbf{B}_t^\top + \mathbf{b}_t\mathbf{b}_t^\top = \mathbf{I}_2$. Rearranging, we have that $\mathbf{B}_t\mathbf{X}\mathbf{X}^\top\mathbf{B}_t^\top - \mathbf{I}_2$ has rank 1 $\forall t$. This leads to the following quartic constraint on $\mathbf{X}$:

$$\det\left(\mathbf{X}\mathbf{X}^\top - \mathbf{K}_t\right) = 0 \tag{5}$$

where $\mathbf{K}_t = \mathbf{B}_t^{-1}\mathbf{B}_t^{-\top}$. Now, denoting the vector of unknowns as $\mathbf{x} = [x_a, x_c, x_d]^\top$ (with $x_b = 0$), given k views we have k constraints from Eq.(5), confirming 3 non-degenerate views are needed for a finite set of solutions [HL89, LF06]. Solving Eq.(5) for $\mathbf{x}$ leads to a 4^{th} order system of 16 monomials, and can easily be solve with Gröbner bases [BJÅ09]. With an estimated solution $\tilde{\mathbf{X}}$, the planar pose in each view is then recovered by factorising each 2×2 sub-block of $\mathbf{B}\tilde{\mathbf{X}}$ using Eq.(2), resulting in the 2-fold solution at each view. The compactness of OAD comes directly from our decoupling euclidean upgrading from ambiguous pose generation, which can then be done for each view independently. In our extensive experiments comprising many thousands of runs, the number of real solutions were found to be between 0 and 2. In the noise free case the correct solution is always given. With noise it is possible for no real solution to exist; since the affine motion has no exact physical interpretation.

Our method can be taken a step further. Given $n \geq 3$ views with noisy measurements we can satisfy Eq.(5) in the least-squares sense by solving for $\mathbf{x}$ such that:

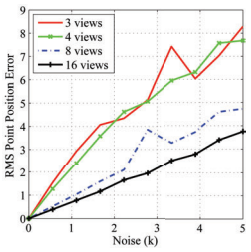
$$\frac{d}{d\mathbf{x}}\sum_{t=1}^{n}\left(\det\left(\mathbf{X}\mathbf{X}^\top - \mathbf{K}_t\right)\right)^2 = 0 \tag{6}$$

However solving Eq.(6) leads to a 3-equation 7^{th} order polynomial system in $\mathbf{x}$. A practical way we can reduce the order is to replace $\mathbf{X}\mathbf{X}^\top$ by the Positive Definite (PD) matrix $\mathbf{W} = \left[\begin{array}{cc} w_1 & w_2 \\ w_2 & w_3\end{array}\right] = \mathbf{X}\mathbf{X}^\top$ and solve for $\mathbf{w} = (w_1, w_2, w_3)$. By relaxing the PD condition on $\mathbf{W}$ we are left with a 3^{rd} order polynomial of 16 monomials. This we again solve efficiently with Gröbner bases. $\tilde{\mathbf{X}}$ may then be recovered from $\tilde{\mathbf{W}}$ via Cholesky decomposition and $\mathbf{B}\tilde{\mathbf{X}}$ can be block factorised as before. In the event that $\tilde{\mathbf{W}}$ is non-PD we currently use the closest least squares PD approximation to $\tilde{\mathbf{W}}$. For $n > 3$ views we have found multiple solutions may be resolved in general by taking the single best solution $\tilde{\mathbf{W}}$ as the one with smallest error (either algebraic from Eq.(5) or reprojection error - see Eq.(7)). Importantly because the number of equations and number of unknowns (*i.e.* 3) do not increase with additional views OAD is practical *for any reasonably large n*.

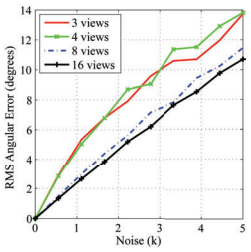
Because OAD minimises an algebraic cost (and is therefore suboptimal in the maximum likelihood sense), pose estimates may be optionally refined via Orthographic Planar Projection Bundle Adjustment (OPP-BA.) If the affine motion has been estimated from point tracks, generated by $p \geq 3$ point samples located on the support plane at unknown positions $\{\mathbf{u}^1, \mathbf{u}^2, .., \mathbf{u}^p\}$, these can be estimated, together with poses $\{\mathbf{R}_t, \hat{\mathbf{t}}_t\}$ optimally from their image correspondences $\{\mathbf{v}_t^i\}$. Assuming these are corrupted by IID gaussian noise this is achieved by minimising the reprojection error:

$$E_{OPP}\left(\mathbf{R}_t, \hat{\mathbf{t}}_t, \mathbf{u}^i; \mathbf{v}_t^i\right) = \sum_{t=1}^{n}\sum_{i=1}^{p}\left(\left[\begin{array}{cc}\hat{\mathbf{R}}_t & \hat{\mathbf{t}}_t \\ \mathbf{0}^T & 1\end{array}\right]\mathbf{u}^i - \mathbf{v}_t^i\right)^2 \tag{7}$$

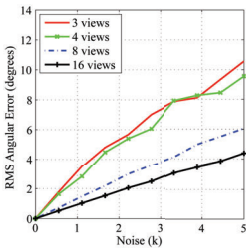
and setting $\mathbf{u}_1 = [0,0]^\top, \mathbf{u}_2 = [0,\cdot]^\top$ to fix the translational and in-plane rotation gauge ambiguities. Once optimised the set $\{\mathbf{u}^i\}$ holds the planar euclidean structure of the point



(a) OAD (tracking): Structure error



(b) OAD (tracking): Pose error



(c) OAD (full measurements): Pose error

	Views / Noise (k)	4		8		12		16	
		OAD	OPP-BA	OAD	OPP-BA	OAD	OPP-BA	OAD	OPP-BA
RMS Angular Error (deg)	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1	4.7	3.8	2.7	1.8	2.6	1.7	2.4	1.6
	2	8.1	7.7	5.1	3.8	4.8	3.2	4.7	3.3
	3	9.9	9.3	7.4	5.6	7.0	4.9	6.8	4.7
	4	11.6	11.7	9.7	7.7	9.0	6.7	8.7	6.4
	5	13.8	14.3	11.5	9.6	11.2	8.4	10.7	7.9
RMS Point Position Error	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1	2.2	2.9	0.9	0.6	0.8	0.4	0.7	0.4
	2	4.3	12.9	1.7	3.1	1.6	0.8	1.5	1.4
	3	5.4	13.2	3.8	3.4	2.4	1.2	2.2	1.0
	4	6.5	22.9	3.9	5.9	3.4	4.1	2.9	1.6
	5	7.7	30.3	4.7	5.9	4.5	3.9	3.8	1.9

(d) OAD vs OPP-BA

Figure 2: *Emperical Performance of OAD*

samples up to a scale factor, and can be used to compute the plane-to-image transforms $\mathbf{A}_t$ from image correspondences.

3.3. OAD: Empirical Results

We now present some synthetic studies to assess the empirical performance of OAD reconstruction; in particular robustness with respect to measurement noise. A planar quad patch was simulated undergoing random rotations, orthographically projected and its four projected corners subjected to additive zero mean Gaussian noise with $SD = \sigma$. For us to present scale independent results, we vary σ relative to the patch's scale: $\sigma = k/100 \times w$ where w denotes the patch's width, set to $w = 100$. First the affine transforms were estimated in a tracking setting: affine transforms between the first and all subsequent views were computed using the corner correspondences, and OAD performed on the affine structure with the least squares formulation (Eq. 6). The corner positions on the support plane were recovered using Eq.(7). This is a linear operation given the image correspondences and recovered plane-to-view transforms. Fig. 2(a) shows the RMS error of the planar point position with respect to k, and Fig. 2(b) the RMS error in the planes' normals. With increased views we observe better robustness to noise, as expected. Interestingly there appears to be little benefit in using 4 views over the 3 view minimal case however. We also tested the performance when a complete measurement matrix $\mathbf{M}$ is provided. This was synthesised by computing interview transforms for all view pairs, with point correspondences subject to varying noise, and the affine structure estimated by taking the rank-2 SVD decomposition of $\mathbf{M}$. Fig. 2(c) shows the performance of the orientation estimates. This marks a clear improvement over Fig. 2(b), becoming substantially better with 16 views, and shows OAD can exploit well the redundancy present in a full measurement matrix. We then investigated the benefits of running OPP-BA in the tracking setting, initialised by the OAD solution. Our results are summarised in 2(d). Our findings indicate that it is in fact detrimental to perform OPP-BA with as few as 4 views at higher noise levels. The benefits only become clearly apparent beyond 8 views. This suggests for some applications the additional cost of running OPP-BA after OAD may not always be worth it.

4. The Weak Template for Planar Pose Disambiguation and Shape Estimation

In §2 we have presented methods for euclidean-upgrading an isolated planar patch using multiple orthographic views. Let us return back to the context of deformable surfaces. For each frame, a surface comprising n tracked planes would result in a 2^n-fold orientation ambiguity. We show in §4.1 and §4.2 this can be reduced, to at best a 2-fold ambiguity by exploiting physical constraints acting between pairs of neighbouring planes. The 2-fold ambiguity corresponds to a global reflection of the surface about the camera's z axis. Treated as independent frames, this ambiguity is irresolvable in orthographic views without additional cues. We propose that with the assumption of temporal continuity, we can recover a unique solution across the video using a disambiguated seed frame. The problem amounts to inferring the template's MRF state for each frame, with energy of the classic form $E(\gamma_t^1, \gamma_t^2, ., \gamma_t^N) = \sum_{(p,q)\in E} \varphi(\gamma_t^p, \gamma_t^q) + \alpha \sum_{i\in V} \phi(\gamma_t^p)$, where $\phi(\gamma_t^p)$ denotes the per-node temporal constraints and $\varphi(\gamma_t^p, \gamma_t^q)$ denotes the pairwise physical constraints with tuning weight α.

4.1. Bending Surface Constraints

What constraints can exist between the poses of two planes located on a deformed surface to solve Necker disambiguation? When these planes are far apart the answer is very little in general. However when in local proximity a local model of surface bending can be used to constrain their poses, and hence be used for disambiguation. Our model uses the fact that inextensible surfaces such as those made from cloth or paper exhibit local developability, and prohibits poses corresponding to high twisting or shearing of the surface.

Formally, nearby tangent planes are constrained by bending about local rulings (Fig. 3.) On the local scale a developable surface is modelled by a parabolic cylinder [MC98], with rulings approximately parallel. Parallel rulings imply that the orientations of two nearby planar patches $\mathcal{P}$ and $\mathcal{Q}$ can be modelled by a hinge system. Fig. 3-(a) illustrates an image

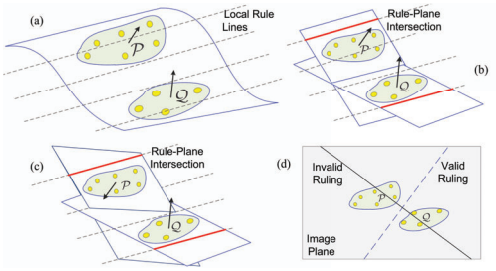


Figure 3: *Surface bending constraining two disjoint patches*

of a locally ruled surface with parallel rule lines shown in dashed, and two planar patches $\mathcal{P}$ and $\mathcal{Q}$. Fig. 3-(b) shows the corresponding hinge system. The red lines indicate the projection of these rulings on the planes supporting $\mathcal{P}$ and $\mathcal{Q}$. The planes are consistent with surface bending if these projections are parallel. Fig. 3-(c) shows a configuration inconsistent with surface bending, and here is caused by the wrong Necker state attributed to $\mathcal{P}$. Note that the bend model does not necessarily correspond to a real physical hinge; it constrains only the orientations of the two patches. Note also that as the surface deforms the rulings may change.

It is possible in theory to recover the rule orientations directly from the planes' affine structures. However estimating rulings from image data is notoriously unstable. Instead we optimise over the range of rule orientations. Suppose we have estimated the 3D orientations of $\mathcal{P}$ and $\mathcal{Q}$ unambiguously from their plane-to-view affine transforms $\mathbf{A}^p$ and $\mathbf{A}^q$ respectively. Call these $\mathbf{R}^p$ and $\mathbf{R}^q$, with normal vectors $\mathbf{n}^p$ and $\mathbf{n}^q$ respectively. Suppose also we have a putative estimate of the angle θ made between the camera's x axis and the 3D rulings. That is, the rulings lie on parallel planes orthogonal to the viewing direction with normals $\mathbf{n}_r = [\cos(\theta), \sin(\theta), 0]^\top$. $\mathcal{P}$ and $\mathcal{Q}$ mutually satisfy the hinge model if the intersection lines between these planes, and $\mathcal{P}$ and $\mathcal{Q}$ are parallel. The model's error is defined as:

$$E(\theta, \mathbf{R}_p, \mathbf{R}_q) = 1/z\,(\mathbf{n}_p \times \mathbf{n}_r) \cdot (\mathbf{n}_q \times \mathbf{n}_r) \tag{8}$$

with $z = \|\mathbf{n}_p \times \mathbf{n}_r\| \, \|\mathbf{n}_p \times \mathbf{n}_r\|$. We optimise θ by sampling over the range $[0 : \pi]$ (we currently use 25 samples), giving the bend error E_b:

$$E_b(\mathbf{R}_p, \mathbf{R}_q) = \underset{0<\theta\leq\pi}{\arg\min} \left[E(\theta, \mathbf{R}_p, \mathbf{R}_q)\right] \tag{9}$$

There exist additional constraints on θ which should be considered, because the extents of $\mathcal{P}$ and $\mathcal{Q}$ constrain the rulings. A rule should not bisect either region, since these are by definition planar (Fig. 3-(d).) Currently, we keep only those θ whose line does not bisect the point samples in $\mathcal{P}$ or $\mathcal{Q}$. If no such angle exists, it implies $\mathcal{P}$ and $\mathcal{Q}$ cannot bend (*i.e.* they are planar in that view) and so we arbitrarily set $\theta = 0$.

4.2. Pose Disambiguation

Now consider when only the planes' affine motion is known, but not their Necker states. The unambiguous orientations are given by $\mathbf{R}^p = [\mathbf{A}^p, \gamma^p]_{3\times3}$ and $\mathbf{R}^q = [\mathbf{A}^q, \gamma^q]_{3\times3}$. The joint 4-fold ambiguity can be brought down by evaluating Eq.(9) using each state permutation, and Necker states violating the hinge model can be detected by inspecting the model error. In fact the ambiguity cannot be totally resolved, but merely brought down to 2-fold. This is a consequence of the Necker reversal of the hinge system itself. It is easy to show that Eq.(9) is of the following form:

$$E_b\left(\left[\mathbf{A}_t^p, \gamma_t^q\right]_{3\times3}, \left[\mathbf{A}_t^q, \gamma_t^q\right]_{3\times3}\right) = \begin{cases} c_1 & \text{if } \gamma_t^p = \gamma_t^q \\ c_2 & \text{otherwise} \end{cases} \tag{10}$$

for some $c_1, c_2 \geq 0$. That is, if we flip the Necker states of both $\mathcal{P}$ and $\mathcal{Q}$ we generate the same fitting error. Returning to the template's MRF, each edge is associated with a symmetric interaction potential derived from Eq.(10). We simply use it directly: $\varphi(\gamma_t^p, \gamma_t^q) = E_b\left(\left[\mathbf{A}_t^p, \gamma_t^q\right]_{3\times3}, \left[\mathbf{A}_t^p, \gamma_t^q\right]_{3\times3}\right)$. In fact there exists surface configurations where the bending model provides no additional constraints. The degeneracy arises when the hinge axis is orthogonal to the camera's z axis where it can be shown that $c_1 = c_2 = 0$ (under perfect modelling conditions.) In these configurations the hinge system provides no constraints. As a result it may be possible for sections of the template to be unconstrained in some frames. Additional constraints are needed.

4.3. Outlier Removal

The bending model can also be violated by outliers; planes with poorly estimated poses due to erroneous affine motion. An outlier plane will usually violate the bending model for most of its edges in the template graph. Given two connecting nodes $\mathcal{Q}$ and $\mathcal{P}$, we deem the edge to violate the model if $\min(c_1, c_2) > 0.35$. $\mathcal{P}$ is marked as an outlier if $r/e \geq 0.8$, where r denotes the number of violating edges.

4.4. Unambiguous Pose with Temporal Continuity

To resolve the global 2-fold ambiguity per frame, and to circumvent the degenerate bend configurations, we can exploit the fact that the surface deforms smoothly over time. The nodes' states in subsequent frames are strongly constrained. This naturally suggests a 3D MRF formulation. However in this paper we opt for a simpler, albeit less optimal strategy: sequentially processing the video and make hard state decisions at previous frames. Suppose at frame t the Necker states of a node have been resolved up to the

$(t-1)^{\text{th}}$ frame. We give preference to its state γ_t^q if the rotation $\left[\mathbf{A}_t^p, \gamma_t^q\right]_{3\times 3}$ is predicted by smooth angular motion. We fit a quaternion smoothing cubic spline (ignoring the unity constraint) to the rotations assigned in the previous $m = 10$ frames. Denote $\tilde{\mathbf{R}}_t^p$ to be the prediction of the spline extrapolate at time t. The MRF's temporal constraints are given by $\phi\left(\gamma_t^p\right) = \left\| \tilde{\mathbf{R}}_t^p - \left[\mathbf{A}_t^p, \gamma_t^p\right]_{3\times 3} \right\|$, with $\|\cdot\|$ being the Frobenius norm. To initialise the temporal constraints, we currently provide a manual disambiguation at frame 1. The MRF contains submodular interaction terms, and so resolving sate is NP-hard. We have however found good success using belief propagation.

4.5. Recovering Nonrigid Shape

After template node disambiguation, we densely reconstruct the deformed surface. Our goal is a 2.5D reconstruction: reconstructing the region $\mathcal{R}_t$ of the deforming surface that is visible in each frame. Formally, we determine at each time the function $S(x,y;\theta_t) : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ that projectively maps image pixels to 3D. In orthographic conditions this is given by:

$$S(x,y;\theta_t) = (x,y,D_t(x,y;\theta)) \quad \forall (x,y) \in \mathcal{R}_t \qquad (11)$$

where $D(x,y;\theta_t)$:$\mathbb{R}^2 \rightarrow \mathbb{R}$ denotes a scalar depth field parameterised by θ_t. Importantly, although the template construction process is run in orthographic conditions, shape estimation may be run in perspective conditions by changing Eq.(11) to a perspective depth function. Planar orientations are estimated in the same way using Eq (2). Recovering depth over $\mathcal{R}_t$ given only a sparse set of orientation estimates is ill-posed (*i.e.* the Poisson equations for normal integration are under-constrained.) If we assume that within $\mathcal{R}_t$ shape is generally smooth this becomes tractable. We cast the reconstruction problem as classic pseudo MAP estimation by minimising the reconstruction energy: $E(S;\theta_t) = E_N(\theta_t) + \lambda E_{bend}(\theta_t)$. The estimated unambiguous normals act as data terms of the form:

$$E_N(\theta_t) = \sum_{\mathbf{p}_i} \left(\left.\frac{\partial S}{\partial x}\right|_{\mathbf{p}_i} \cdot \mathbf{n}_i \right)^2 + \left(\left.\frac{\partial S}{\partial y}\right|_{\mathbf{p}_i} \cdot \mathbf{n}_i \right)^2 \qquad (12)$$

where $\mathbf{p}_i$ denotes the locations of the point samples in the image with normals $\mathbf{n}_i$ within each inlier plane. For the smoothing term E_{bend} we use the TPS regulariser penalising the second order bend energy of $S(x,y;\theta_t)$. This is weighted by λ that currently we experimentally set. The system is solved by quantising $\mathcal{R}_t$ with a dense quad mesh. θ_t now holds the mesh vertices' depths. We use a finite differences to approximate the surface tangents in Eq.(12) [NRDR05] and TPS bend energy [PLF05]. These are linear operators, leading to a sparse linear system in θ_t, and thus solved efficiently.

4.6. Experimental Results and Practical Considerations

In this section we show some results our NR-SfM approach applied to two real sequences. The first sequence is of a deforming creased sheet of paper with printed text (Fig. 4.) This comprises 80 frames taken in approximate orthographic conditions. Frames 1 and 80 shown in Fig. 4-(a,e). A ROI $\mathcal{R}$ was manually marked in frame 1 (shown in blue) and keypoints within $\mathcal{R}$ were tracked using KLT, and clustered into affine groups (shown in Fig. 4-(b).) Each colour denotes a cluster, with white denoting an outlier point track not assigned to a cluster. The weak template was constructed with nodes corresponding to clusters and edges taken from a Delaunay triangulation of the clusters' centres in the first frame, and keep only the edges contained within $\mathcal{R}$. OAD was then performed on each node. In Fig. 4-(f) we show the projection of the unambiguous normal in frame 1. Normals coloured in blue denote detected outliers. $\mathcal{R}$ was then transferred throughout the sequence by warping the ROI in the first frame using affine moving least squares [SMW06]. In Fig. 4-(c,g) we show the reconstructed surfaces at frames 1 and 80. Qualitatively the results look convincing, particularly in capturing the crease edge. To inspect the reconstruction's quality, we flattened the surface at frame 1 onto the 2D plane (shown in Fig. 4-(d)). The results suggest a faithful reconstruction, highlighting applications for monocular document restoration. Finally in Fig. 4-(h) we transferred a different texture to the 3D surface, showing the reconstruction was sufficiently good for augmented reality. Next we processed the sequence used in [SUF08] (Fig. 5), comprising 87 frames of a bending cardboard surface. This is quite challenging for templateless reconstruction because of the texture sparsity. The affine point clusters are shown in Fig. 5-(a). We show the projection of the unambiguous normals in Fig. 5-(b,c,d) at frames 9, 27, 45 and 65 respectively. The corresponding surface reconstructions are shown below each image rendered from a different view and the normals and reconstructions appear faithful. However with no ground truth data available quantitative performance results are unavailable.

5. Conclusion and Future Work

We have presented new methods for solving NR-SFM using the assumption of local planarity and rigidity. Solutions have been given for planar structure and motion in orthographic conditions, called Orthographic Affine Decomposition (OAD). This provides closed form solutions to the minimal 3-view and general $n > 3$-view cases. Secondly, we have proposed the idea of a weak deformable template for surface reconstruction; a surface abstraction with nodes holding local planar structure and edges corresponding to pairwise physical constraints embodying a local bending model. In conjunction with temporal continuity, the ambiguities can be brought down to a unique solution across a video sequence. As future work we aim to make disambiguation fully automatic and perform fuller quantitative performance analysis of the 3D reconstructions for more complex scenes. We wish to extend the scope of our work to handle scenes with self occlusions, handle lost point tracks and ultimately reconstruct complete 3D surfaces from partial reconstructions.

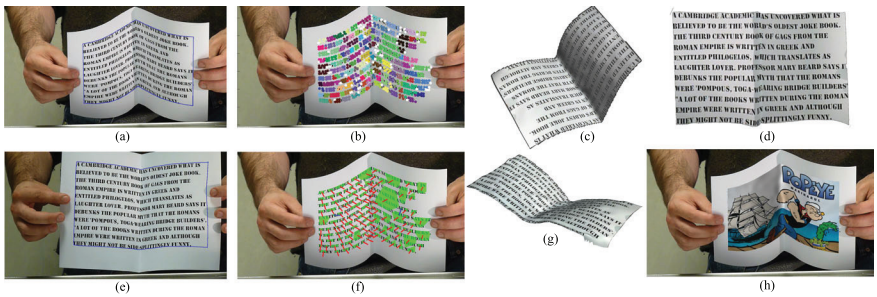


Figure 4: *Reconstruction of creased paper*

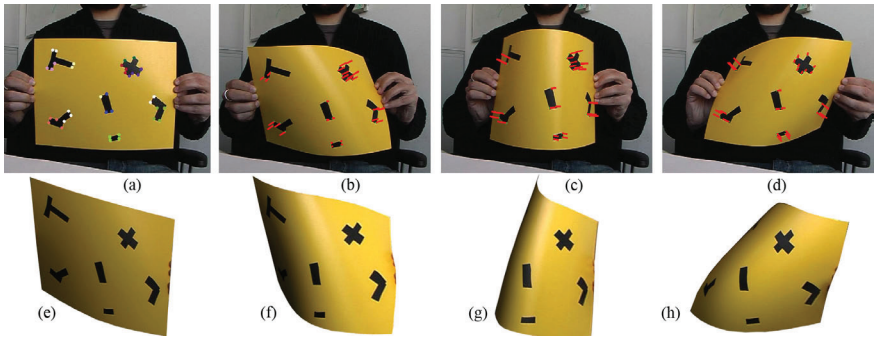


Figure 5: *Reconstruction of a sparsely textured surface*

References

[BHB*10] BRUNET F., HARTLEY R., BARTOLI A., NAVAB N., MALGOUYRES R.: Monocular template-based reconstruction of smooth and inextensible surfaces. In *ACCV* (2010). 1

[BJÅ09] BYRÖD M., JOSEPHSON K., ÅSTRÖM K.: Fast and stable polynomial equation solving and its application to computer vision. *IJCV* (2009). 4

[CDGB10] COLLINS T., DUROU J.-D., GURDJOS P., BARTOLI A.: Single-view perspective shape-from-texture with focal length estimation: A piecewise affine approach. In *3DPVT* (2010). 1, 3

[FXC09] FERREIRA R., XAVIER J., COSTEIRA J.: Shape from motion of nonrigid objects: The case of isometrically deformable flat surfaces. In *BMVC* (2009). 1, 2

[HL89] HUANG T. S., LEE C. H.: Motion and structure from orthographic projections. *PAMI* (1989). 4

[LF06] LOBAY A., FORSYTH D. A.: Shape from texture without boundaries. *IJCV* (2006). 1, 2, 4

[MC98] MAEKAWA T., CHALFANT J.: Computation of inflection lines and geodesics on developable surfaces. In *Mathematical Engineering in Industry* (1998). 6

[NRDR05] NEHAB D., RUSINKIEWICZ S., DAVIS J., RAMAMOORTHI R.: Efficiently combining positions and normals for precise 3d geometry. *TOG* (2005). 7

[PHB10] PERRIOLLAT M., HARTLEY R., BARTOLI A.: Monocular template-based reconstruction of inextensible surfaces. *IJCV* (2010). 1

[PLF05] PILET J., LEPETIT V., FUA P.: Real-time non-rigid surface detection. In *CVPR* (2005). 7

[SF09] SALZMANN M., FUA P.: Reconstructing sharply folding surfaces: A convex formulation. In *CVPR* (2009). 1

[SMW06] SCHAEFER S., MCPHAIL T., WARREN J.: Image deformation using moving least squares. *TOG* (2006). 7

[SSL10] SHEN S., SHI W., LIU Y.: Monocular 3d tracking of inextensible deformable surfaces under l2-norm. *Trans. Img. Proc.* (2010). 1

[SUF08] SALZMANN M., URTASUN R., FUA P.: Local deformation models for monocular 3d shape recovery. In *CVPR* (2008). 1, 7

[TJK10] TAYLOR J., JEPSON A., KUTULAKOS K.: Structure from locally-rigid motion. In *CVPR* (2010). 2

[VSTF09] VAROL A., SALZMANN M., TOLA E., FUA P.: Template-free monocular reconstruction of deformable surfaces. In *ICCV* (2009). 1

[WF06] WHITE R., FORSYTH D. A.: Combining Cues: Shape from Shading and Texture. In *CVPR* (2006). 2