Template-free Shape from Texture with Perspective Cameras

Anna Hilsmann[12]
anna.hilsmann@hhi.fraunhofer.de
David C. Schneider[12]
david.schneider@hhi.fraunhofer.de
Peter Eisert[12]
peter.eisert@hhi.fraunhofer.de

[1] Computer Vision & Graphics Group
Fraunhofer Heinrich Hertz Institute
Berlin, Germany

[2] Visual Computing Group
Humboldt University of Berlin
Berlin, Germany

The term Shape-From-Texture (SFT) covers a class of methods for computing the 3D shape of a textured surface from a single image by exploiting texture distortion as a cue for shape. In this paper, we present an SFT formulation, which is equivalent to a *single-plane/ multiple-view* pose estimation problem statement under perspective projection. As in the classical SFT setting, we assume that the texture is constructed of one or more repeating texture elements, called *texels*, and assume that these texels are small enough such that they can be modeled as planar patches. In contrast to the classical setting, we do not assume that a fronto-parallel view of the texture element is known a priori. Instead, we formulate the SFT problem akin to a Structure-From-Motion (SFM) problem, given n views of the same planar texture patch. In contrast to many SFT methods that use orthographic or scaled orthographic camera models [1, 2, 4], we assume a full perspective camera model with known intrinsics and estimate the patch poses from estimated homographies between the distorted texel appearances in the image. Our method is inspired by recent work of Varol *et al.* [6] who used homography decomposition [5, 7] to reconstruct deforming surfaces in monocular video sequences.

Let $\mathbf{P}_0 = \mathbf{K}\left[\mathbf{I}|\mathbf{0}\right]$ and $\mathbf{P}_1 = \mathbf{K}\left[\mathbf{R}|\mathbf{t}\right]$ be the projection matrices of two cameras with the same intrinsics $\mathbf{K}$ and the rigid motion between them described by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. Assume a plane with normal vector $\mathbf{n}$ and distance d from the origin is projected into both camera frames. The projections of a point on the plane into the images of $\mathbf{P}_0$ and $\mathbf{P}_1$ are then related by a homography which is given as [3]

$$\mathbf{H} = \mathbf{R} - \frac{\mathbf{t}\mathbf{n}^T}{d} \quad . \tag{1}$$

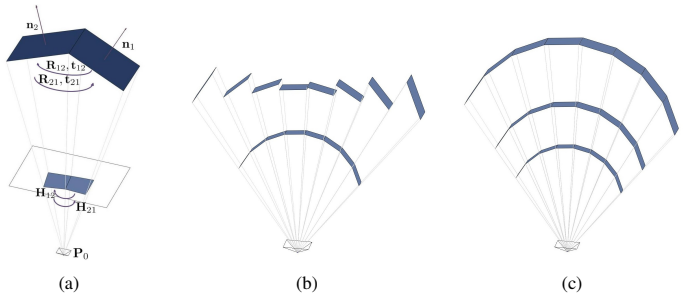
Methods for decomposing a homography into $\mathbf{R}, \mathbf{t}/d$ and $\mathbf{n}$ are detailed in [5, 7]. The decomposition results in two distinct solutions for the (scaled) relative (camera) motion and the plane normal in the camera referential coordinate frame. This ambiguity can be solved if a third view of the plane is given as one solution of each decomposition yields the same normal vector for the reference plane.
The above considerations are equivalent to the assumption of a fixed camera $\mathbf{P}_0$ and a *moving* plane. In the SFT setting, the image contains n views of the same planar patch or, equivalently, n identical planar patches, i.e. the texture elements, under rigid motion. Let $\mathbf{H}_{kl}$ denote the homography between two patches k and l. The decomposition of each homography $\mathbf{H}_{kl}$ yields an estimate of the (scaled) rigid motion from the reference patch k to patch l and of the reference patch normal

$$\mathbf{H}_{kl} \Rightarrow \mathbf{R}_{kl},\ \mathbf{t}_{kl}/d_k,\ \mathbf{n}_{k_l}$$

where $\mathbf{n}_{k_l}$ denotes the l^{th} estimate of $\mathbf{n}_k$ (see figure 1(a) for an illustration with two planes). Note that the translation vector contains a scale ambiguity. By using each patch as reference to all other patches in turn, we get enough constraints to minimize a large sparse linear least squares cost function optimizing the 3D poses of the texel patches. The cost function is based on the estimated rigid motion between planes and the reprojection error as well as an additional planarity constraint. For a single point i on patch k the cost function yields:

$$\begin{aligned}\hat{\mathbf{X}}_k^i = \arg\min_{\mathbf{X}_k^i} \sum_{l=1}^{n} \left\| \left[\mathbf{R}_{kl}|\mathbf{t}'_{kl}\right] \cdot \begin{bmatrix}\mathbf{X}_k^i \\ 1\end{bmatrix} \times \begin{bmatrix}\mathbf{x}_l^i \\ 1\end{bmatrix} \right\|^2 \\ + \sum_{l=1}^{n} \left\| {\mathbf{n}_{k_l}}^T \cdot \mathbf{X}_k^i + d_k \right\|^2 \quad .\end{aligned} \tag{2}$$

with $\mathbf{R}_{kk} = \mathbf{I}$ and $\mathbf{t}_{kk} = \mathbf{0}$. To reconstruct all points on all patches this results in a large linear equation system. The local scale ambiguity (figure 1(b)) in the rigid motion estimation can either be solved jointly or separately by providing additional constraints based on the rigid motions between corresponding 3D points. Detailed derivations are given in the paper.
A smooth surface is computed by regression with approximating thin-plate splines using the estimated patch centroids as data points. The final reconstruction is up to a single global scale factor (figure 1(c)).



(a) (b) (c)

Figure 1: (a) Scenario with two planes, (b) local scale ambiguity, (c) global scale ambiguity

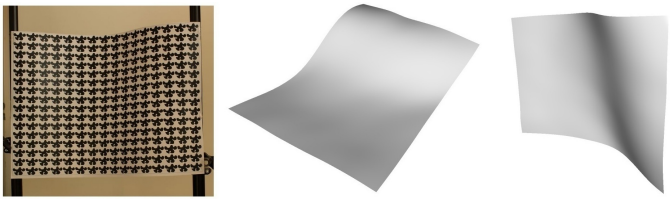


Figure 2: Input image (left) and reconstructed surface.

Synthetic experiments were performed to asses the reconstruction quality with respect to noise and different viewing conditions (e.g. with increasing distance between camera and object perspective distortions get weaker and viewing conditions tend to more affine conditions). The reconstruction quality was measured using the angular RMSE of the surface normals. Experiments on real data (figure 2) were evaluated by comparison of the interpolated surface with a reconstruction from stereo correspondences which showed a mean angular RMSE of the surface normals of 4.9 degrees.

[1] T. Collins, J.-D. Durou, A. Bartoli, and P. Gurdjos. Single-View Perspective Shape-from-texture with Focal Length Estimation: A Piecewise Affine Approach. In *Proc. 3D Data Processing, Visualization and Transmission (3DPVT 2010)*, 2010.

[2] D. A. Forsyth. Shape from Texture without Boundaries. In *Proc. Europ. Conf. on Computer Vision (ECCV 2002)*, pages 225–239, 2002.

[3] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[4] A. Lobay and D. A. Forsyth. Recovering Shape and Irradiance Maps from Rich Dense Texton Fields. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 1, pages 400–406, Los Alamitos, CA, USA, 2004.

[5] E. Malis and M. Vargas. Deeper Understanding of the Homography Decomposition for Vision-Based Control. Arobas INIRA Sophia Antipolis, Universidad der Sevilla, 2007. Technical report.

[6] A. Varol, M. Salzmann, E. Tola, and P. Fua. Template-Free Monocular Reconstruction of Deformable Surfaces. In *Proc. Int. Conference on Computer Vision (ICCV 2009)*, 2009.

[7] Z. Zhang and A. R. Hanson. Scaled Euclidean 3D Reconstruction Based On Externally Uncalibrated Cameras. In *In IEEE Symposium on Computer Vision*, pages 37–42, 1995.