# Selecting Surface Features for Accurate Multi-Camera Surface Reconstruction

Thomas Popham
tpopham@dcs.warwick.ac.uk
Roland Wilson
rgw@dcs.warwick.ac.uk

Signal and Image Processing Group
Department of Computer Science
University of Warwick
Coventry, CV4 7AL, England

Establishing correspondences between different views of an object is a longstanding problem in computer vision and is a fundamental requirement in many computer vision applications. One approach to the problem is to detect features in every view and then find correspondences by matching robust feature descriptors [3, 6]. A second approach is to detect features in a reference image and then use standard stereo reconstruction techniques to find the correspondences in the other images [1, 4]. In the second approach, the key requirement is that it will select the best textures for surface reconstruction. This paper addresses this requirement, by asking the question: how should local image textures be selected, so that the visible surface can be accurately reconstructed using stereo techniques?

The contribution of this paper is a novel feature detector that extracts image textures for which a planar patch can be accurately fitted to the corresponding scene surface. We focus on planar patch fitting techniques [1, 5], as these offer a higher level of generality than techniques which assume the surface surface is parallel to the image plane. The proposed feature detector is based upon finding textures that are sensitive to shear transformations, which are part of the chain of projective transformations between two cameras via a plane [2].

Three parameters must be determined to fit an image patch to the scene surface: one for the depth $d$ and two for describing the surface orientation $\theta_1, \theta_2$. An accurately fitted patch has the property that when the texture from each camera is projected onto the patch, the texture remains fixed. Therefore patches are fitted by minimising a cost function, which is the difference between an image patch in a reference camera and the image from second camera which is projected into the reference camera via the plane. The transformation of image co-ordinates from one camera to another via a plane is called an homography, which is a 3x3 matrix with eight degrees of freedom. The problem of fitting a patch to the scene surface may therefore be formulated as one of finding the homography which minimises the reprojection error. Using the notation $x = [d, \theta_1, \theta_2]$ to describe the state vector of the patch, the reprojection error between a window $w_j$ in camera $j$ and the corresponding image pixels in camera $j$ is:

$$\varepsilon^2_{i\to j}(x) = \sum_{g_j \in w_j} \left(z_j(g_j) - z_i(H_{ji}(x) g_j)\right)^2 \tag{1}$$

where $z_j(g_j)$ is the image intensity of the homogenous point $g_j$ in camera $j$ and $H_{ji}(x)$ is the homography between camera $j$ and camera $i$, which is parameterised by the plane vector $x$. Where multiple cameras are available, the reprojection error is summed over the set of multiple cameras $\mathscr{C}$:

$$\varepsilon^2(x) = \sum_{j \in \mathscr{C}} \varepsilon^2_{i\to j}(x) \tag{2}$$

In order to find the parameters of the planar patch, a natural requirement is that only the true plane parameters $x$ minimise the objective function in equation (2). With multiple cameras, the depth of the feature is usually sufficiently constrained, but accurate surface normal estimation is often difficult, as different orientations of the plane may give rise to similar reprojection errors. Checking that every combination of $\theta_1$ and $\theta_2$ produces a unique texture would be a computationally intensive method of ensuring accurate surface orientation estimation. The proposed method therefore checks the sensitivity of the local texture to an affine transformation, and in particular a shear transformation is used. Both x-shear and y-shear transformations are parameterised as follows:

$$H_{sx}(c_x) = \begin{bmatrix} 1 & c_x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad H_{sy}(c_y) = \begin{bmatrix} 1 & 0 & 0 \\ c_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

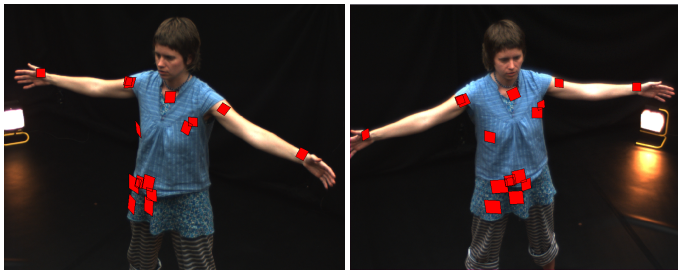

Figure 1: Two views of estimated patches from proposed feature detector

where both $H_{sx}$ and $H_{sy}$ are transformations of a point $(x,y,1)^T$ and $c_x$,$c_y$ control the amount of shear in the x and y directions. The sensitivities $r_x(u,v)$ and $r_y(u,v)$ of the texture window $f$ to the transformations are defined as:

$$r_x(u,v) = \left(\frac{\partial f}{\partial c_x}\right)^2 \qquad r_y(u,v) = \left(\frac{\partial f}{\partial c_y}\right)^2 \tag{4}$$

and these gradients are estimated by comparing an original texture window $f(m,n)$ against a small transformation of the texture:

$$r_x(u,v) = \sum_{(m,n)} \left(f(m+u,n+v) - f(m+u+c_x n, n+v)\right)^2 \tag{5}$$

$$r_y(u,v) = \sum_{(m,n)} \left(f(m+u,n+v) - f(m+u, n+v+c_y m)\right)^2 \tag{6}$$

Since two parameters must be estimated for the surface orientation, we want the texture to be responsive to shear transformations in both the x and y directions. The actual response $a(u,v)$ at each pixel is therefore the square-root of the product of the two responses:

$$a(u,v) = \sqrt{r_x(u,v) r_y(u,v)} \tag{7}$$

The paper describes two experiments. The first experiment compares the performance of the proposed detector with a SIFT keypoint extractor [3], using a synthetic dataset. This experiment shows that patches originating from the proposed detector are more accurately fitted to the scene surface than patches originating from the SIFT keypoint extractor. The second experiment shows a real-world application of the proposed detector, using a performance capture dataset. Figure 1 shows the patches overlaid onto the two different views of the scene, and it is clear that the patches are fitted close to the scene surface.

[1] M. Habbecke and L. Kobbelt. A Surface-Growing Approach to Multi-View Stereo Reconstruction. In *Proc. CVPR*, pages 1–8, 2007.

[2] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[3] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[4] H.P. Moravec. *Robot spatial perception by stereoscopic vision and 3D evidence grids*. Carnegie Mellon University, The Robotics Institute, 1996.

[5] A. Mullins, A. Bowen, R. Wilson, and N. Rajpoot. Multiresolution particle filters in image processing. In *Proceedings of the Mathematics in Signal Processing Conference*, 2006.

[6] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.