Robust Estimation of Multiple Surface Shapes from Occluded Textures

Michael J. Black
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
black@parc.xerox.com

Ruth Rosenholtz
Helmholtz Instituut, Utrecht University
Princetonplein 5, 3584 CC Utrecht
the Netherlands
R.Rosenholtz@fys.ruu.nl

Abstract

This paper examines the problem of estimating surface shape from texture in situations in which there are multiple textures present due to texture discontinuities, occlusion, and pseudo-transparency (for example looking through a picket fence at a textured surface). Previous shape-from-texture methods that use changes in the spatial frequency representation of neighboring image patches assume that only a single texture is present in each of the patches. We extend these approaches to situations in which multiple textures may be present. We provide a theoretical analysis of the multiple texture problem and the effect of texture discontinuities, occlusion, etc. on the spatial frequency representation. We also present an algorithm, using robust mixture models, for recovering multiple surface shapes from occluded textures. The method performs well on real and synthetic images with results which are comparable to that of shape from texture with only one texture.

1 Introduction

In this paper we explore the problem of estimating surface shape from texture in situations in which there are multiple textures present in a region of interest. Multiple textures due to texture discontinuities, occlusion, and pseudo-transparency pose problems that are closely related to the analogous problems in motion and stereo. Unlike recent work in stereo and motion, however, shape-from-texture methods typically assume that only a single texture is present in an area of interest. We relax this *single texture assumption* and exploit techniques from the robust estimation of multiple motions to recover multiple surface shape estimates in the presence of multiple textured surfaces.

Figure 1 shows examples multiple textures that occur in natural scenes; they are of four types:

(a) Texture Outlier: An image patch contains a brightness structure that is not consistent with the dominant texture.

(b) Texture Discontinuity: Multiple textures within a single image region.

(c) Fragmented Occlusion or Pseudo-Transparency: One surface is viewed *through* another surface such as a fence.

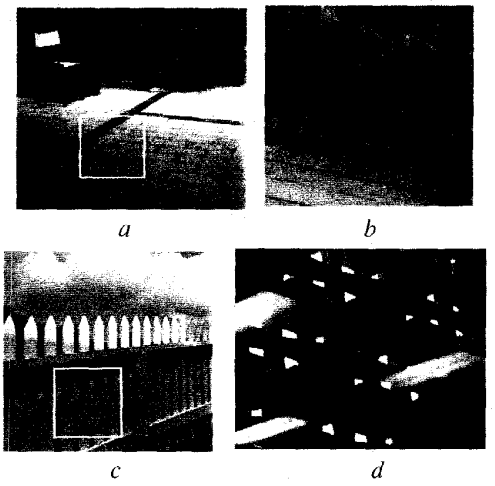


Figure 1: Multiple textured surfaces with occlusion.

(d) Transparent/Reflected Textures: eg. viewing a textured surface through glass which is also reflecting a textured surface. Without additional cues such as motion, images of this sort can be confusing to human observers. We leave this problem for future work.

We take as our basic model the shape-from-texture methods based on local distortions of the spatial frequency of the texture [5, 6, 7, 8]. Given the spatial frequency representation of a texture in an image patch, we look at neighboring image patches and, in frequency domain, compute the affine texture distortion from one image patch to another. In certain situations when multiple textures are present, the spatial frequency representation contains a (possibly weighted) sum of the significant frequency components corresponding to each of the different textures. We refer to multiple textures of this form as "additive" in the frequency domain.

In this additive case, a spectrogram containing peaks from multiple textures can be thought of as consisting of a number of "layers" where each layer corresponds to the spatial frequencies present in a single texture (see Figure 2). Our goal is to compute a set of "weights" that assign spatial frequencies to layers and to estimate the affine transformation for each layer. We also want the estimated transformations to be robust to spurious peaks and noise in the spectro-

0-8186-7190-4/95 $4.00

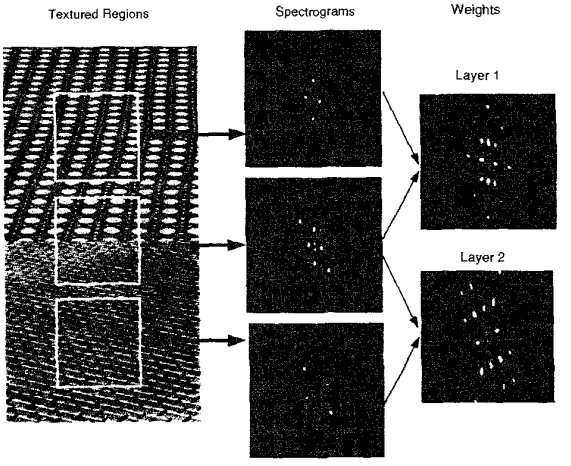


Figure 2: Additive texture split into layers.

grams. When textures are additive in the frequency domain we can apply tools developed for dealing with multiple motions to the new problem of multiple textures.

Until recently, work in motion estimation has exploited a *single motion assumption* which states that within an arbitrary image region only a single image motion is present. This assumption is violated in common situations such as when the region contains transparency or occlusion. To relax this assumption, a number of approaches represent the motion of an image region as a set of layers, where the motion of each layer is described by a low-order parametric model (eg. affine) [1, 2, 4, 9]. In particular, Jepson and Black [4] use an iterative algorithm (the EM-algorithm) to assign optical flow constraints to a set of layers and estimate the affine motion of each layer. In this paper we apply a similar approach to the problem of estimating the affine texture distortions in frequency domain.

2 Shape-from-Texture

We begin with the basic approach of Malik and Rosenholtz [8] which uses a modified version of windowed Fourier transform magnitude [7] as a measure of the statistics of the texture in an image region. They model the local *texture distortion* as a set of affine transforms, A_i, between spectral estimates of an image patch and eight neighboring patches. They derived the relationship between these affine transforms and the local shape and orientation of the surface. The surface parameters of interest are the slant, σ, the direction of tilt, $\mathbf{t}$, (θ_t is the angle between this vector in the image plane and the x-axis), and the curvature parameters, $r\kappa_t$, $r\kappa_b$, and $r\tau$. Slant is the angle between the surface normal vector and the line of sight. Here κ_t and κ_b are the normal curvatures of the surface in the tilt direction ($\mathbf{t}$) and the direction perpendicular to the tilt direction ($\mathbf{b}$), respectively. The variable r is the distance from the center of the viewing sphere to the given point on the surface, and τ is the "geodesic torsion" of the surface in the tilt direction. For the mathematical relationship between the local affine transformations and the surface shape parameters see [8].

They use a differential method for finding these affine transforms which resembles differential methods used to compute optical flow. They assume that one spectrogram differs from another by only a 2×2 affine transformation $A = I + \Delta A$, so $F_2(\vec{\omega}) = F_1(A\vec{\omega})$, where F_i is the spectrogram for a patch centered about the ith point, and $\vec{\omega}$ is the frequency. Then, for small ΔA, we can write

$$F_2(\vec{\omega}) - F_1(\vec{\omega}) \approx \nabla\vec{F}_1 \circ \Delta A\vec{\omega} \tag{1}$$

where $\nabla\vec{F}_1$ is the gradient of the spectrogram at the given frequency. Malik and Rosenholtz solve for ΔA using least squares. We will replace their estimation technique with a robust technique that can estimate multiple transformations simultaneously.

Finally, Malik and Rosenholtz use non-linear minimization of the error between the empirical affine transformations and the theoretical affine transformations (for a given shape estimate) to solve for shape and orientation estimates:

$$\chi^2(\sigma, \theta_t, r\kappa_t, r\kappa_b, r\tau) = \sum_{i=1}^{n}\sum_{k=1}^{2}\sum_{l=1}^{2}(\hat{A}_i(k,l) - \bar{A}_i(k,l))^2$$

where $\hat{A}_i(k,l)$ is the (k,l)th element of the theoretically predicted matrix $\hat{A}_i$ and is a function of the shape parameters, and $\bar{A}_i(k,l)$ is the (k,l)th element of the empirically measured affine transform matrix $\bar{A}_i$.

3 Motivation and Theoretical Analysis

The canonical example for shape from texture for multiple textures with occlusion is the fragmented occlusion example shown in Figure 1*c*. We will refer to fragmented occlusion using terminology suggested by the example shown of a wood fence in front of a grassy field, though of course the analysis applies in general when one texture partially occludes another. In space domain, we have

$$\begin{aligned}\textbf{image} \quad = \quad & \textbf{wood texture} \times \textbf{where the fence is} \\ & + \textbf{grass texture} \times \textbf{where the fence isn't}\end{aligned}$$

where **wood texture** is the texture actually on the boards of the fence and **where the fence is/isn't** are binary functions indicating the location of the boards of the fence and the spaces between them, respectively. Transforming into frequency domain, and letting W, G, F and $\overline{\mathrm{F}}$ be the Fourier transforms of **wood texture**, **grass texture**, **where the fence is** and **where the fence isn't**, respectively, we get

$$F(\textbf{image}) = \mathrm{W} * \mathrm{F} + \mathrm{G} * \overline{\mathrm{F}}. \tag{2}$$

At the scale shown in the figure, one really wants to know the orientation and shape of the fence and the orientation and

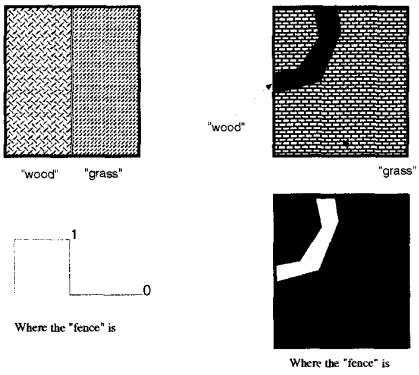


Figure 3: Texture with discontinuity (left) and with occlusion or outliers (right) can be thought of as special cases of the fragmented occlusion problem.

shape of the grassy field, but not the shape and orientation of the individual boards that make up the fence. In this case, we are interested in when Equation 2 can be approximated by the "additive" equation

$$F(\mathbf{image}) = k_1 \mathrm{F} + k_2 \mathrm{G} \tag{3}$$

for some constants, k_1 and k_2.

Note that the cases of texture discontinuity and shape from texture with occlusion are basically subsets of this canonical example of fragmented occlusion as depicted in Figure 3.

3.1 When is the "Additive" Approximation Valid?

We need a model of the later processing in the Fourier domain in order to define what we mean by a valid approximation. We will assume that all frequency components that have magnitude which is N or more times lower than the magnitude of the largest frequency component are suppressed. For this analysis, we will assume that the textures are periodic, so that they have well-defined peaks in frequency domain at frequencies ω_i, for some i.

Fragmented Occlusion. To simplify the situation somewhat, we will assume that the wood texture, W, can be approximated by only a dc component. In other words, we assume that the boards of the fence are all untextured or at least that any "wood grain" texture on the boards is relatively low power.

Then, splitting $\overline{\mathrm{F}}$ into its dc component and the remaining frequency components, we can rewrite Equation 2 as

$$F(\mathbf{image}) \;=\; k_1 \mathrm{F}(\omega) + k_2 \mathrm{G}(\omega) + \sum_{\omega_i \neq 0} k_i \mathrm{G}_i(\omega)$$

where k_1 and k_2 are constants representing the dc values of W and $\overline{\mathrm{F}}$, respectively. The first two terms are a weighted sum of F and G, as desired, but that we also have additional copies of the grass texture, G, convolved with the non-dc components of $\overline{\mathrm{F}}$. We have written these additional copies as $k_i \mathrm{G}_i(\omega)$. We not only have extra copies of the grass texture, but those extra copies effectively undergo different affine transforms because their affine transforms are centered about the frequencies ω_i.

Without loss of generality, we assume that the highest peak in frequency domain corresponds to the F texture. We refer to the dc component of any given texture, T, as $T(0)$. Then we write the highest peak in T, not counting the dc component, as $T(1)$, the next highest peak $T(2)$, and so on.

In order for the frequency components to be separable into two textures two conditions must be met. First the extra copies of the grass texture, G must be small enough that they are suppressed; i.e. the frequency components are lower than $1/N$ times the highest frequency component. This can be expressed as the condition that

$$|\mathrm{W}(0) - \mathrm{G}(0)|/N > |\mathrm{G}(1)|. \tag{4}$$

Intuitively, this first condition states that the contrast between the wooden fence and the grass ($|W(0) - G(0)|$) must be high enough compared to the contrast of the grass texture ($|G(1)|$).

The second condition requires that we see enough frequency components in each texture to be able to estimate shape from texture. We must not suppress either the F texture or the first copy of the G texture and there must be at least n peaks in the first copy of both textures in order to be able to perform shape from texture on both textures. If we are not to suppress the first copy of the grass texture, we must have the following condition:

$$|[\mathrm{W}(0) - \mathrm{G}(0)]\mathrm{F}(1)|/N < |\mathrm{G}(n)\overline{\mathrm{F}}(0)|. \tag{5}$$

where $|\mathrm{G}(n)\overline{\mathrm{F}}(0)|$ is the magnitude of the nth highest peak associated with the first copy of the grass texture.

Intuitively, $\overline{\mathrm{F}}(0)$ is the percentage of pixels in the image in which one can see the grass texture. Therefore, for the fence example, this second constraint is satisfied if there is enough space between the boards of the fence for us to see "enough" of the grass to recover shape from the grass texture.

Texture Discontinuities and Outliers The cases of shape from texture in the presence of a discontinuity of texture and in the presence of an occluder turn out to be much simpler. This is because both the discontinuity and the occluder tend to be low frequency; otherwise we would have the previous case of fragmented occlusion. Therefore, F and $\overline{\mathrm{F}}$ are each a single low frequency "blob" in frequency domain, and, Equation 2 becomes a blurred version of the desired $k_1 \mathrm{W} + k_2 \mathrm{G}$. When this additive approximation holds we should be able to separate the two textures in frequency domain and perform the shape from texture on both textured surfaces.

4 Estimating Multiple Affine Transformations

With a single texture we can robustly recover the affine transformation, ΔA, between neighboring spectrograms using the differential method described in [7]. Taking Equation (1) as a constraint, we solve for the affine transformation ΔA between spectrograms F_2 and F_1 by minimizing

$$\sum_{\vec{\omega}} \rho(\nabla\vec{F}_1 \circ \Delta A\vec{\omega} - (F_2(\vec{\omega}) - F_1(\vec{\omega})), \sigma). \quad (6)$$

where ρ is some robust error norm; for example

$$\rho(x, \sigma) = \frac{x^2}{\sigma + x^2} \quad (7)$$

which is used in [1]. As the magnitudes of residuals $\nabla\vec{F}_1(\vec{\omega})^T \circ \Delta A\vec{\omega} - (F_2(\vec{\omega}) - F_1(\vec{\omega}))$ grow beyond a point their influence on the solution begins to decrease and the value of $\rho(\cdot)$ approaches a constant while the weight goes to zero.

To estimate the affine transformations for multiple surfaces we must do two things: (1) we must decide which frequencies $\vec{\omega}$ correspond to which texture, and (2) given this correspondence estimate the transformation for each texture. This first step can be thought of as assigning frequencies to a set of layers corresponding to the textures. To estimate multiple affine transformations ΔA_i we minimize

$$\sum_{\vec{\omega}} m_{\vec{\omega},i}(\nabla\vec{F}_1(\vec{\omega})^T \circ \Delta A_i\vec{\omega} - (F_2(\vec{\omega}) - F_1(\vec{\omega})))^2 \quad (8)$$

for each ΔA_i where $m_{\vec{\omega},i}$ is a weight that encodes the likelihood that the frequency $\vec{\omega}$ belongs to the layer i. The weights are defined in terms of a robust error norm

$$m = \frac{1}{2x}\frac{\partial}{\partial x}\rho(x, \sigma) = \frac{\psi(x, \sigma)}{2x} = \frac{\sigma}{(\sigma + x^2)^2}. \quad (9)$$

The objective function (8) is minimized in two stages. First we solve for the weights, $m_{\vec{\omega},i}$, in closed form using Equation (9)

$$m_{\vec{\omega},i} = \frac{\sigma}{(\sigma + (\nabla\vec{F}_1(\vec{\omega})^T \circ \Delta A_i\vec{\omega} - (F_2(\vec{\omega}) - F_1(\vec{\omega})))^2)^2}.$$

After updating all the weights, we can update the estimates of ΔA_i using weighted least squares or gradient descent.

A frequency $\vec{\omega}$ may be shared by multiple layers, but if there are no outliers present, each frequency must belong to some layer. This can be expressed as the following *mixture constraint*

$$\sum_i m_{\vec{\omega},i} = 1. \quad (10)$$

Dealing with Outliers. The mixture constraint assumes that every frequency belongs to some layer or is shared by multiple layers. It does not account for outliers which belong to no layer. To cope with outliers, we follow the approach in [3] and introduce a new *outlier layer* that does not correspond to any affine transformation and is not used in the minimization of Equation (8) but is used when enforcing the mixture constraint. If we are estimating n affine transformations then the normalization in Equation (10) is computed over the $n+1$ layers. The initial value for the $m_{\vec{\omega},n+1}$ is taken to be the weight given to the largest expected outlier.

The values of $m_{\vec{\omega},n+1}$ are updated only when normalizing the weights. If a frequency $\vec{\omega}$ does not correspond to any affine transformation ΔA_i then it will receive low weights and the normalization will shift weight to the outlier layer and $m_{\vec{\omega},n+1}$ will increase.

Implementation. In our current implementation we assume that the number of layers is known. To estimate the affine transformation of each layer we first construct a Gaussian pyramid representation of the spectrograms under consideration. The affine transformation is computed at a coarse level and then, at the next finer level, the estimated transformation is used to register the two patches by warping one towards the other. This process is repeated down to the finest level in the pyramid while the transformation is updated at each stage.

Within each level of the pyramid the affine transformation is estimated using a gradient descent method embedded within a continuation strategy in which the value of σ begins at a high value and is gradually lowered [1]. The effect of this process is to gradually reduce the influence of frequencies which are treated as outliers with respect to one or all of the layers.

Recovering Multiple Shapes. In the case of a single texture, Malik and Rosenholtz [7] recover surface shape by computing the transformations between a single patch and eight neighboring patches. In the case of multiple textures, each pair of image patches gives rise to n affine transformations. To estimate multiple shapes from these multiple transformations we need to know which transformations correspond to which of the n surfaces. We group the individual layers recovered from each of the eight neighboring patches into a single set of layers using a simple match metric that takes into account where the weights of the layers agree and where they disagree. Once the correspondence between the layers of the different patches is known, we use the shape estimation procedure described in Section 2 with the affine transformations corresponding to each set of layers to recover the multiple surface shapes.

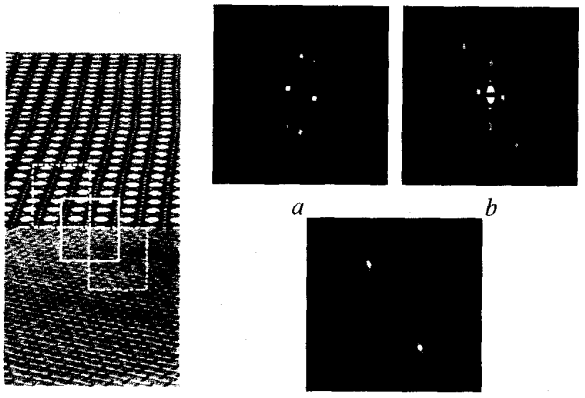


Figure 4: Texture Discontinuity. Left: Image showing example patch locations. Right, Spectrograms: *(a)* upper left patch. *(b)* center left patch. *(c)* lower right patch.

Figure 5: Discontinuity example. Weights estimated for two of the eight patches.

5 Results

We present a number of results on natural and synthetic images that indicate the accuracy of the recovered surface shapes are comparable to those reported by Malik and Rosenholtz [8] for shape from texture with only one texture.

5.1 Texture Discontinuities

Figure 4 illustrates a synthetic texture discontinuity. Nine patches (128 $\times$ 128 pixels) are extracted with the central patch centered on the texture boundary. The neighboring patches are offset from the center by 64 pixels. A two level Gaussian pyramid was used with 40 iterations of the continuation method at each level. The value of σ began at 70.0 and was lowered by a factor of 0.9 after each iteration to a minimum of 35.0. Peaks with a height lower than 63.0 were ignored.

Figure 5 shows the recovered weights for the patches (*a* and *c*) in Figure 4. The two patterns corresponding to the different textures are clearly visible. Using the recovered affine transformations between the center patch and each of the eight neighboring patches we recover the following surface shape estimates for the top and bottom surfaces:

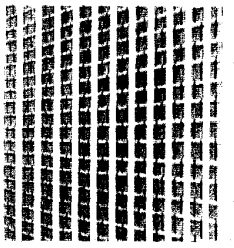


Figure 6: Fragmented Occlusion. Both the center region and the surrounding regions contain multiple textures.

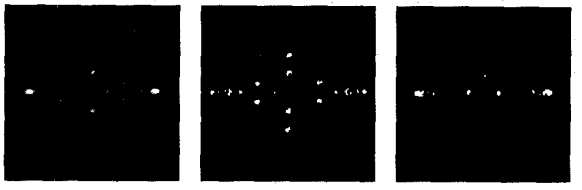


Figure 7: Fence Example. *(a)* Center spectrogram. Normalized weights corresponding to: *(b)* the background (grass). *(c)* the foreground (fence).

Surface	slant	tilt	rk_t	rk_b	$r\tau$
Top (True)	70	-90	0	0	0
Bottom (True)	70	-90	0	0	0
Top (Est.)	65.4	-86.7	-0.52	-0.16	-0.01
Bottom (Est.)	60.0	-93.7	-0.04	-0.06	-0.11

The slight underestimates of slant are due to the fact that the algorithm assumes that the step in the image was from the center of one patch to the center of another. The effective step, due to the discontinuity, is between the center of the *texture* in one patch to the center of the same texture in the other patch.

5.2 Synthetic Fence Example

The next experiment considers the fragmented occlusion case in Figure 6 in which we have a "fence" pattern in front of a textured surface (the "grass"). The patches are taken to be 128 $\times$ 128 with the neighboring patches offset from the center patch by 32 pixels. A four level Gaussian pyramid was used with 20 iterations of the continuation method at each level. The value of σ began at 60.0 and was lowered by a factor of 0.95 after each iteration to a minimum of 20.0. Peaks with a height lower than 90.0 were ignored.

Figure 7 shows the averaged weights from all the layers corresponding to one or the other surface. Notice that the power in the "fence" texture lies along a line in frequency domain. This means that the estimation of the affine transformations for the fence texture will be underconstrained. This has been referred to as the texture aperture effect [8].

The true and estimated parameters for the fence example are:

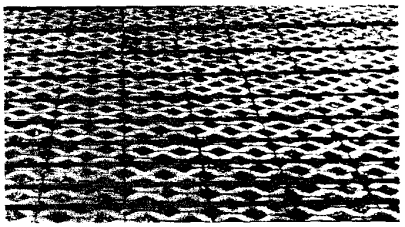
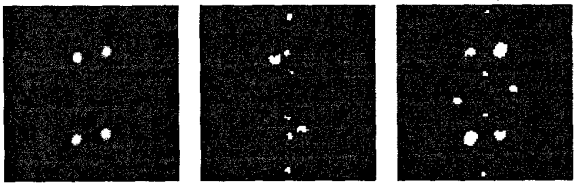


Figure 8: Natural image of a ground plane viewed through Venetian blinds.



a *b* *c*

Figure 9: Venetian Blind Example. *(a)* Center spectrogram. Normalized weights corresponding to: *(b)* the foreground (blinds). *(c)* the background (ground plane).

Surface	slant	tilt	rk_t	rk_b	$r\tau$
"Grass" (True)	70	-180	0	0	0
"Fence" (True)	0	undef	0	0	0
"Grass" (Est.)	67.9	-185.4	-0.15	-0.05	-0.04
"Fence" (Est.)	8.7	38.6	-7.14	-4.54	16.7

Notice that the shape of the background textured surface has been recovered accurately. For the fence, the slant of 8.7 is fairly accurate while the true tilt is undefined for a slant of 0 degrees. The high curvature terms are due to the aperture effect. In general one should detect when the solution is ill-conditioned due to the texture aperture effect and indicate which results are reliable and which are not. A solution to this problem is beyond the scope of this paper.

5.3 Seeing Through Blinds

We now consider a natural scene exhibiting fragmented occlusion. Figure 8 shows a textured ground plane (the "grass") viewed through Venetian blinds (the "fence"). The patches are taken to be 128 × 128 with the neighboring patches offset from the center patch by 64 pixels. Exactly the same parameters were used as were used in the previous synthetic example.

Figure 9 shows the spectrograms and the recovered weights which are averaged across the corresponding layers of all the patches. Notice that the blinds, once again, suffer from the texture aperture problem while the peaks corresponding to the ground-plane texture are clearly recovered.

The recovered surface parameters are:

	Estimated				
Surface	slant	tilt	rk_t	rk_b	$r\tau$
"Grass"	58.3	-87.4	0.02	0.13	0.09
"Fence"	16.9	-147.7	2.28	2.38	10.99

While we do not have ground truth for this image, the "grass" texture appears to accurately reflect the orientation of the ground plane. Likewise the slant of the blinds appears to be well recovered but due to the aperture problem we cannot hope to recover the full orientation and shape accurately.

6 Conclusions

Previous work in shape-from-texture assumes that, within the region of interest, there is only one texture present, and no occluders. But it is quite common in natural scenes to see one textured surface through another, as in viewing our canonical example of a grassy field through a fence. We have shown theoretically that for a large class of problems, the shape from texture problem for multiple textures and occlusion is analogous to robust estimation of structure from motion. We have presented a robust shape from texture method based on a combination of Malik and Rosenholtz' [8] affine model of shape from texture and previous work in robust affine motion estimation [1, 4], and have demonstrated results on both synthetic and natural images which achieve the accuracy of methods that estimate shape from texture with only one texture.

References

[1] M. Black and P. Anandan. The robust estimation of multiple motions: Affine and piecewise-smooth flow fields. Tech. Rep. P93-00104, Xerox PARC, Dec. 1993.

[2] T. Darrell and A. Pentland. Robust estimation of a multi-layer motion representation. In *IEEE Workshop on Visual Motion*, pp. 173–178, Princeton, Oct. 1991.

[3] S. Gold, C. P. Lu, A. Rangarajan, S. Pappu, and E. Mjolsness. Fast algorithms for 2D and 3D point matching: Pose estimation and correspondence. Tech. Rep. YALEU/DCS/RR–1035, Yale University, May 1994.

[4] A. Jepson and M. J. Black. Mixture models for optical flow computation. In *CVPR-93*, pp. 760–761, NY, June 1993.

[5] J. Krumm and S. Shafer. Shape from periodic texture using the spectrogram. In *CVPR'92*, pp. 284–301, Champaign-Urbana, 1992.

[6] J. Krumm and S. Shafer. Segmenting textured 3D surfaces using the space/frequency representation. *Spatial Vision*, Vol. 8, No. 2, pp. 281–308, 1994.

[7] J. Malik and R. Rosenholtz. A differential method for computing local shape-from-texture for planar and curved surfaces. In *CVPR-93*, pp. 267–273, NY, June 1993.

[8] J. Malik and R. Rosenholtz. Recovering surface curvature and orientation from texture distortion: A least squares algorithm and sensitivity analysis. *ECCV-94*, pp. 353–364, Stockholm, 1994.

[9] J. Y. A. Wang and E. H. Adelson. Layered representation for motion analysis. In *CVPR-93*, pp. 361–366, NY, June 1993.