Affine height landscapes for monocular mobile robot obstacle avoidance.

Bojian Liang, Nick Pears and Zezhi Chen
Department of Computer Science, University of York, UK

Abstract. In this paper, we propose a method to segment the ground plane from a mobile robot's visual field of view and then measure the height of non-ground plane features (even raw pixels) above the mobile robot's ground plane. Thus a mobile robot can determine what it can drive over, what it can drive under, and what it needs to manoeuvre around. In addition to obstacle avoidance, this data could also be used for localisation and map building. All of this is possible from an uncalibrated camera (raw pixel coordinates only), but is restricted to (near) pure translation motion of the camera. The main contributions are (i) a novel reciprocal-polar (RP) image rectification, (ii) ground plane segmentation by sinusoidal model fitting in RP-space and (iii) a novel projective construction for measuring affine height.

1 Introduction

In this paper, we focus on the application of mobile robot (uncalibrated, monocular) obstacle avoidance and we present a system that can construct an affine height landscape of the robots visible (indoor) environment. The height recovered is termed an *affine height* as it is a height ratio (affine invariant property), referenced to the height of the camera optical centre above the ground plane. The term *landscape* is used as the other two dimensions are view-based (i.e. untransformed pixel coordinates). Affine height (h_a) measurements allow potential obstacles to be classified as either small enough to be driven over (we require $h_a < 0.1$), high enough to be driven under (we require $h_a > 1.25$), or true obstacles to be avoided.

We present three main new results: First, by expressing images in a reciprocal-polar $(\frac{1}{r}, \alpha)$ form with the origin on the focus of expansion (FOE) [1], the image motion of a set of co-planar points along the $\frac{1}{r}$ direction is a pure shift, when the translation is parallel to the plane. This allows image motion to be accurately recovered by 1D correlation, even over large image distortions caused by large camera motion. Second, we show that the magnitude of these shifts follows a sinusoidal form along the α direction over a maximum of π radians. Simultaneous ground plane pixel grouping and recovery of the ground plane homography thus amounts to finding the FOE and then robustly fitting a sinusoid, whose phase corresponds to the orientation of the vanishing line of the ground plane and whose amplitude is related to the magnitude of the robot/camera translation. The method allows every ground plane pixel in a locally textured region to contribute to the estimation of the ground plane homography, thus giving a highly accurate result. Finally, our third new result shows that, given the homography associated with the ground plane, the affine height of remaining non-ground plane pixels, referenced

[1] The FOE is the point where the direction of the translation vector, passing through the camera optical centre, intersects the image plane. It is where all image motion emanates from under pure camera translation

to the height of the camera optical centre above the ground plane, can be determined using the virtual parallax cue computed using a construction based on the cross-ratio.

Our algorithms require camera motion that is (near) pure translation. Obviously, it can be argued that this is restrictive, but such motions are common, can be deliberate in mobile robot applications and can easily be detected over an image pair, particularly when corner correspondences are available. Also, given that an affine height landscape has been computed, it can be tracked through robot motions which have a rotational component and new unlabelled areas of the scene that enter the field of view can be probed by further translation motions.

In the following sections, we first discuss ground plane motion (and hence homography) recovery, which simultaneously gives a ground plane segmentation. We then show how the recovered homography can be used to measure affine height.

2 Ground plane segmentation and ground plane motion/homography recovery

Early work on exploiting coplanar relations has been presented by Tsai and Huang [5], Longuet-Higgins [2] and Faugeras and Lustman [4]. We summarise the coplanar relation as follows: If a set of feature points in the scene lie in a plane, and they are imaged from two viewpoints, then the corresponding points in the two images are related by a planar homography, H, such that $\lambda \mathbf{x_j} = \mathbf{H}\mathbf{x_i}$, where $\mathbf{x}$ represents a homogenous image coordinate $(x, y, 1)^T$, $\mathbf{H}$ is a 3 by 3 matrix representing the homography and λ is a scalar. Since this equation is valid up to a scale factor, $\mathbf{H}$ has only eight degrees of freedom.

Suppose that a mobile robot (and therefore camera) attempts to move under pure translation. Due to an uneven floor surface and hysteresis in the robot's mechanics, the motion is unlikely to be pure translation. However, if rotation is relatively small with respect to the translation, assuming pure translation and enforcing a homography corresponding to pure translation allows correlation based techniques to be used. The key point here is that this allows *all* ground plane pixels which have local intensity/color variation to be used in the simultaneous estimation of the ground plane homography and grouping of ground plane pixels.

Under pure translation, a planar homography is often termed a planar homology [3], which has five degrees of freedom (dof) and takes the form

$$\mathbf{H} = \mathbf{I} - k\mathbf{x}_f\mathbf{l}_v^T \tag{1}$$

where $\mathbf{x}_f = [x_f, y_f, 1]^T$ is the focus of expansion (FOE) for the two frames, $\mathbf{l}_v = [a_v, b_v, 1]^T$ is the vanishing line (or horizon line) of the plane and k is a constant associated with the magnitude of the translation.

Firstly, we check if (near) pure translation is detected, by intersecting all lines defined by all corner correspondences from the image pair. If most lie in a small area (95% of intersections lie should lie within a circle of small radius), then translation is assumed and the FOE is computed using random sample consensus (RANSAC) and least squares (LS). Once the FOE has been computed, we shift image coordinates so that each image is centered on the FOE. Thus we apply the centering translation $\mathbf{T}_c$ where:

$$\mathbf{T}_c = \begin{bmatrix} 1 & 0 & -x_f \\ 0 & 1 & -y_f \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

After this translation, the FOE is at homogenous coordinates $\mathbf{x}'_f = [0, 0, 1]^T$ and the vanishing line becomes $\mathbf{l}'_v = \mathbf{T}_c^{-T}\mathbf{l}_v = [a_v, b_v, \mathbf{x}_f^T\mathbf{l}_v]^T$. Thus, the homography relating points in FOE centered coordinates is $\mathbf{H}' = \mathbf{I} - k\mathbf{x}'_f\mathbf{l}_v^{T\prime}$, and substituting this in the equation $\lambda\mathbf{x}'_2 = \mathbf{H}x'_1$ and expanding gives

$$\lambda \begin{bmatrix} x'_2 \\ y'_2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -ka_v & -kb_v & (1-kq) \end{bmatrix} \begin{bmatrix} x'_1 \\ y'_1 \\ 1 \end{bmatrix} \tag{3}$$

where $q = \mathbf{x}_f^T\mathbf{l}_v$. We note that for translation parallel to the ground plane, $q = 0$ since the FOE lies on the vanishing line. In this specialisation, the homography has four dof and is sometimes called an elation [3]. Otherwise, the FOE is at a distance, $d = \frac{q}{\sqrt{a_v^2+b_v^2}}$ from the vanishing line. To simplify notation, we now drop the 'prime' notation from equation 3 and assume that (x, y) are image measurements made relative to the FOE. Thus, we have

$$(-ka_vx_1 - kb_vy_1 + 1 - kq)x_2 = x_1 \tag{4}$$

$$(-ka_vx_1 - kb_vy_1 + 1 - kq)y_2 = y_1 \tag{5}$$

Squaring both sides and adding

$$(-k(a_vx_1 + b_vy_1) + (1 - kq))^2(x_2^2 + y_2^2) = (x_1^2 + y_1^2) \tag{6}$$

If we define r_i as the Euclidean distance between an image point and the FOE in frame i, then, taking square roots of equation 6

$$(-k(a_vx_1 + b_vy_1) + (1 - kq))r_2 = r_1 \tag{7}$$

$$r_2^{-1} = -k(a_vx_1 + b_vy_1)r_1^{-1} + (1 - kq)r_1^{-1} \tag{8}$$

$$r_2^{-1} = -k(a_v\cos\alpha + b_v\sin\alpha) + (1 - kq)r_1^{-1} \tag{9}$$

where α is the angular position of a pixel in a frame centred on the FOE. Now the gradient of the vanishing line is given as $\tan\alpha_v = \frac{-a_v}{b_v}$, so

$$a_v = -\sqrt{(a_v^2 + b_v^2)}\sin\alpha_v, \quad b_v = \sqrt{(a_v^2 + b_v^2)}\cos\alpha_v \tag{10}$$

Hence

$$r_2^{-1} = k_p(-\sin\alpha_v\cos\alpha + \cos\alpha_v\sin\alpha) + k_qr_1^{-1} \tag{11}$$

where

$$k_p = -k\sqrt{(a_v^2 + b_v^2)}, \quad k_q = 1 - kq \tag{12}$$

Hence, we arrive at the key equation, which defines a function of the angle α, $s(\alpha)$, as

$$s(\alpha) = \rho_2 - k_q\rho_1 = k_p\sin(\alpha - \alpha_v) \tag{13}$$

where we define $\rho = \frac{1}{r}$. Equation 13 indicates that we need to find three constants (k_q, k_p, α_v) in order to recover the homography and that the computation should be implemented in (ρ, α) image space. (Note that a planar homology has five dof, but two have been recovered in the FOE computation.) We call $I(\rho, \alpha)$ reciprocal-polar (RP) image space. Thus, after computing the FOE, an interpolation procedure is used to generate a (possibly scaled) RP image for each image in the image pair.

For the set of planes parallel to the translation direction (which includes the ground plane), the expected value of k_q will be unity, as the expected value of q is zero. For certain applications (hard flat floor, hard robot wheels) it may be reasonable to assume k_q is unity, *as we do*, in other applications it may be preferable to estimate this value, although any estimated value is likely to be very close to unity. For each pixel in image 1, its position in RP image space is computed, and a 1D window is created around this position along the ρ dimension. We then correlate this window along the ρ dimension in RP image 2, at the same value of α. The position of the maximum value of the correlation is retained as a value of $s_i(\alpha)$. Equation 13 indicates an important result: *coplanar motions in RP image space lie on a sinusoid and the constants* (k_p, α_v) *may be recovered by fitting a sinusoid to the RP motion data,* $s(\alpha)$. Suppose that we have two values of s, $s_{i,j}$ measured at two angles, $\alpha_{i,j}$, so that

$$s_i = k_p \sin(\alpha_i - \alpha_v), \quad s_j = k_p \sin(\alpha_j - \alpha_v) \tag{14}$$

hence

$$\frac{s_i}{s_j} = \frac{\sin\alpha_i \cos\alpha_v - \cos\alpha_i \sin\alpha_v}{\sin\alpha_j \cos\alpha_v - \cos\alpha_j \sin\alpha_v} \tag{15}$$

$$\frac{s_i}{s_j} = \frac{\sin\alpha_i - \cos\alpha_i \tan\alpha_v}{\sin\alpha_j - \cos\alpha_j \tan\alpha_v} \tag{16}$$

collecting terms in $\tan\alpha_v$ and rearranging gives

$$\tan\alpha_v = \frac{s_j \sin\alpha_i - s_i \sin\alpha_j}{s_j \cos\alpha_i - s_i \cos\alpha_j} \tag{17}$$

Thus a pair of s values, at different angular positions, for pixels belonging to the same plane, allows us to estimate the orientation of the vanishing line of that plane. Given the phase angle, α_v, corresponding to the orientation of the vanishing line, we can compute k_p from equation 14. By selecting random pairs of angles, $\alpha_{i,j}$ and computing the associated magnitude and phase of the sinusoid upon which both s_i and s_j lie, a random sample consensus (RANSAC) procedure [1] can be used to determine the best set of inliers in the $s(\alpha)$ data to a putative sinusoid. This putative sinusoid is used to initialise an iterative procedure where a LS estimate of the sinusoid parameters and the associated set of inliers are computed until the inlier distribution, represented by a binary tag string, stabilises or the maximum number of iterations is reached. In this way, co-planar pixels may be grouped without explicit construction of a homography matrix. Now, let

$$s_i = k_p \sin(\alpha_i - \alpha_v) = m \sin\alpha_i - n \cos\alpha_i \tag{18}$$

where, $m = k_p \cos\alpha_v = -kb_v, \quad n = k_p \sin\alpha_v = -ka_v$. Thus, for the inliers of the sinusoid model, we can write

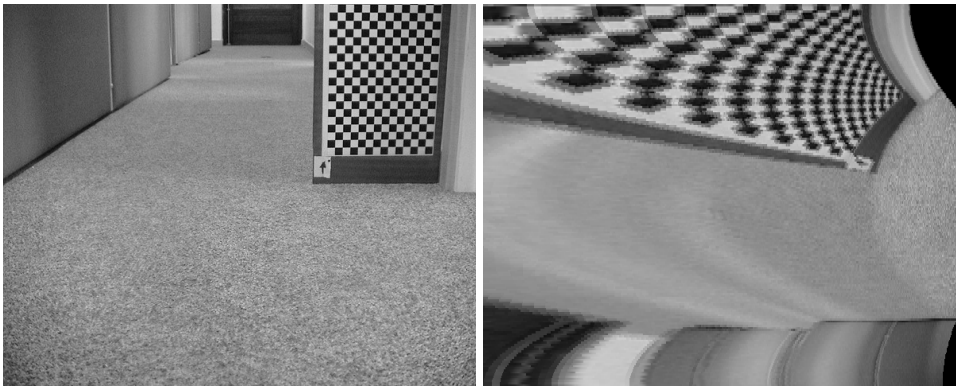

Fig 1a. Original image, $I(x, y)$.

Fig 1b. RP image, $I(\rho, \alpha)$.

$$\begin{bmatrix} -\sin\alpha_i & \cos\alpha_i & s_i \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} m \\ n \\ 1 \end{bmatrix} = \mathbf{0} \tag{19}$$

We use SVD to solve for $\lambda[m, n, 1]^T$ and normalise the solution to obtain the parameters, (m, n). From equation 3, we substitute $n = -ka_v$, $m = -kb_v$ and $k_q = 1 - kq$, to recover the homography in a FOE centred frame. The parameters defining $s(\alpha)$ can be computed as $k_p = \sqrt{(m^2 + n^2)},\ \alpha_v = \tan^{-1}\frac{-n}{m}$ and the homography in the original image frames can be computed as $\mathbf{H} = \mathbf{T}_c^{-1}\mathbf{H}'\mathbf{T}_c$.

How do we know that the recovered homography and grouped pixels are associated with the ground plane? A weak assumption regarding the pose of the camera with respect to the ground plane, suggests that the sinusoid phase should be close to zero (near horizontal vanishing line). Also, since translation is roughly parallel to this plane, the FOE should lie very close to the vanishing line.

To test whether we can recover the sinusoidal model suggested by the analysis, $k_q = 1$ was assumed and two frames were captured before and after the robot moved in the translation mode. The images were then converted using an interpolation process to RP (ρ, α) form. Fig. 1 shows one of the original images (a) and its RP transform (b) with α rendered in the vertical direction and ρ in the horizontal direction. For each pixel in the original image (2nd of pair), we find its position in RP space, and find the maximum correlation value by correlating along a line of constant α (i.e horizontally) in the first RP image of the pair. (We map the second to the first rather than vice-versa, due to field of view considerations). The maximum value of correlation is retained as a value for $s(\alpha)$. The plot of $s(\alpha)$ against α is shown in fig 2(a) for all pixels. Those pixels motions that correspond to the ground plane can clearly be seen to lie on a sinusoid and ground plane pixels can be segmented as inliers to this sinusoid and

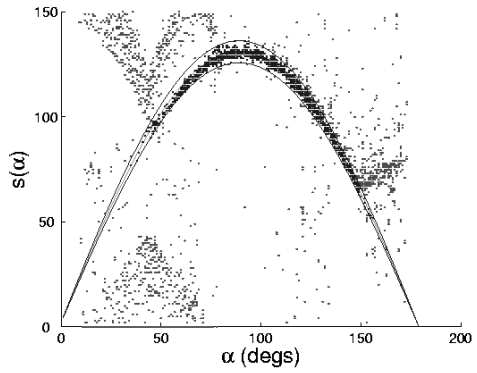

Fig 2a. RP image motion.

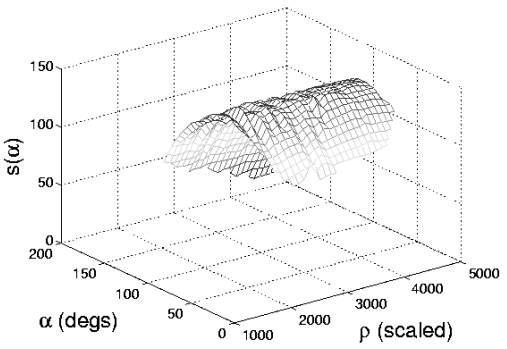

Fig 2b. Ground plane RP image motion.

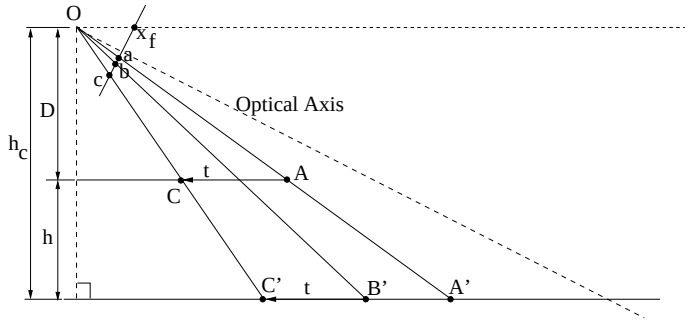

Fig 3 Computation of affine height of point A.

used to compute the ground plane homography. The inliers of the recovered sinusoid are then plotted in 3D in fig 2(b), with the third dimension representing ρ. This indicates that the same sinusoidal form captures image motion in RP-space, irrespective of a pixel's distance from the FOE.

3 Affine height measurement

The approach described above allows pixels to be classified as either belonging to the ground plane or not. For those non-ground plane regions, we would like to know whether we can drive over/under them or whether they form part of an obstacle which should be avoided. We now develop a method of affine height measurement, referenced to the height of the camera optical centre above the ground plane.

Our aim is to recover the height of feature point A shown in figure 3, when the robot undergoes pure (forward) translation, t (and thus the scene point translates t units towards the robot). Point A is the actual position of the feature point relative to the camera before the translation and point C is the position of the feature after the translation. Points A' and C' are the projections of these actual feature positions onto the ground plane. Points a and c are the image positions of the feature at positions A and C respectively and b is the predicted image position of the feature point, if the feature point were to lie in the ground plane. Image point b is computed from the recovered homography induced by the ground plane as: $\mathbf{b} = \mathbf{Ha}$. Now the height of the feature point relative to the height of the camera optical centre, the affine height, is

$$h_a = \frac{h}{h_c} = 1 - \frac{D}{h_c} \tag{20}$$

Using similar triangles, and denoting the distance between points x and y as $d(x, y)$, we note that:

$$\frac{D}{h_c} = \frac{d(OC)}{d(OC')} = \frac{d(AC)}{d(A'C')} \tag{21}$$

For pure translation, $d(A, C) = d(A', B')$, so that

$$h_a = 1 - \frac{d(A'B')}{d(A'C')} \tag{22}$$

Now, the four image points (a, b, c, x_f), where x_f is the focus of expansion (FOE), and the corresponding four ground plane points (A', B', C', ∞) are collinear. The cross ratio for this set of points remains invariant under projection and so we can write:

$$\frac{d(A'B')}{d(A', C')} = \frac{d(a, b)\, d(c, x_f)}{d(a, c)\, d(b, x_f)} \tag{23}$$

Hence, for features below the vanishing line, we can compute affine height as:

$$h_a = 1 - \frac{d(a, b)\, d(c, x_f)}{d(a, c)\, d(b, x_f)} \tag{24}$$

In general, we find that

$$h_a = 1 + \mu \frac{d(a, b)\, d(c, x_f)}{d(a, c)\, d(b, x_f)} \tag{25}$$

where $\mu = -1$ for features below the vanishing line and $\mu = +1$ for features above the vanishing line. (Obviously $h_a = 1$ for features on the vanishing line.) This can be interpreted as the height of point A in units of height h_c. Note that this approach only needs the ground plane homography, H, and the image correspondences a and c of the feature to determine the height above the ground plane. By thresholding the measured height above the plane, the method can be used to check for areas which can be driven over, and for sufficiently high features which can be driven under. Note that this is achieved without camera calibration.

To validate our method of height measurement, a chess board, which has squares of dimension equal to 20% of the height of the camera optical center, was placed in the front of the vehicle, such that it is perpendicular to the ground plane. The vehicle moved forward in pure translation mode and the homography associated with the ground plane was recovered.

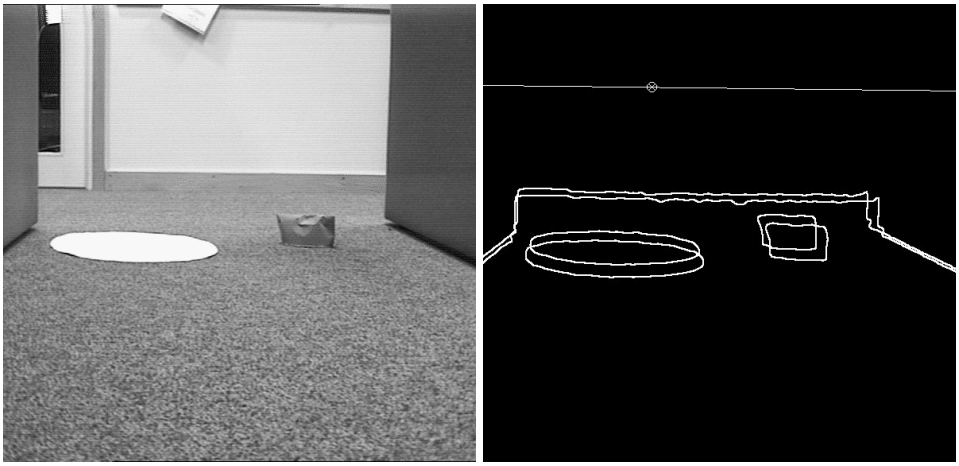

Fig 4a. Raw Image. Fig 4b. Extracted region boundaries.

The affine heights of the corners extracted on the chess board perpendicular to the ground are listed in table 1 for frame 21. In this frame, the image of a chess board square is less than 5% of the total image height. The columns are marked from left to right and the rows from top to the bottom. Ideally, the measurement in each row should be equal and should increase by 0.2 units of the camera height from bottom to top. Note that, due to errors in the measured corner positions and the radial distortion near the image boundary, a number of measurements errors are observed. An error analysis is too detailed to present here, however, these initial results give us confidence that the approach is accurate enough to allow reliable decisions to be made about navigable regions.

Table 1: Affine height of corners on the chess board.

	Column							
	1	2	3	4	5	6	7	8
1	1.67	1.67	1.65	1.68	1.67	1.67	1.67	x
2	1.49	1.50	1.48	1.50	1.52	1.48	1.48	1.48
3	1.34	1.31	1.32	1.32	1.32	x	1.31	1.33
4	1.14	1.15	1.14	1.15	1.14	1.14	1.14	1.13
5	0.97	0.97	0.97	x	0.97	0.96	x	0.97
6	0.82	0.81	0.82	0.79	0.80	0.79	0.76	0.76
7	0.66	0.62	0.63	0.63	0.61	0.60	0.60	0.60
8	x	0.43	0.41	0.43	0.43	0.42	0.42	0.39
9	0.28	0.23	0.20	0.23	0.26	0.24	0.22	0.23
10	x	0.09	0.05	0.08	0.08	0.03	0.09	0.04

The final experiment presented in this paper, uses both the sinusoid fitting process (to simultaneously recover the textured regions of the ground plane and the associated ground

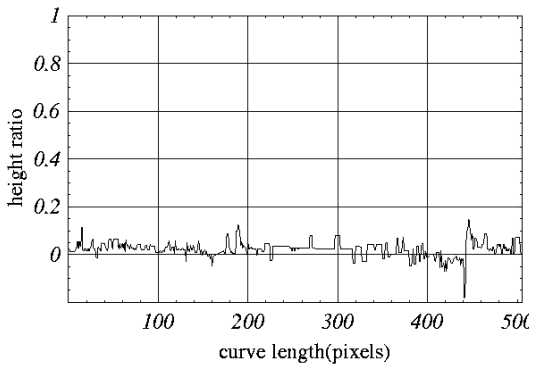

Fig 5a. Height profile of white paper.

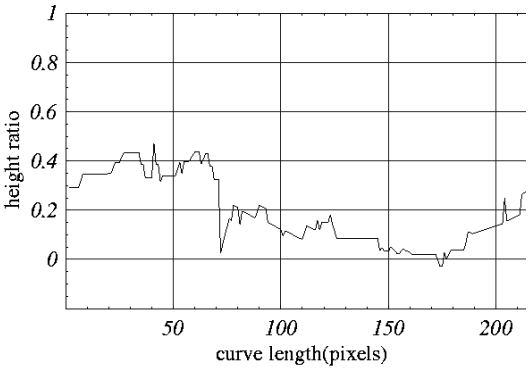

Fig 5b. Height profile of small box.

plane homography) *and* the affine height measurement method to determine whether the contours of *non-textured* regions belong to the ground plane or not. Note that an additional process is required, not described in this paper, which is a quadtree split-merge region segmentation algorithm, which extracts homogenous regions of colour-texture. Textureless regions cannot be classified as ground plane or non-ground plane as they cannot be matched across an image pair. Their boundaries, however, can be and, in the case of pure translation, this is easily done using the vanishing point recovered in the homography estimation process.

Fig. 4(a) shows an image with two regions on the floor which have little texture. The first is a circular piece of white paper, which can be driven over, and the second is a small cardboard box, which can't. Fig. 4(b) shows the extracted boundaries and the vanishing point used to cast a ray in order to match intersections between corresponding boundaries. The cross-ratio construct to measure affine height is applied to the correspondences, thus allowing a height profile to be extracted as we 'walk around' the closed contours associated with the two low texture regions. If the height profile remains close to zero, then the region can be classified as belonging to the ground plane, as in fig. 5(a). Otherwise it is classified as an obstacle, as in fig. 5b. The final image, fig. 6 shows the extracted ground region, where the textured carpet has been classified on a pixel by pixel basis, and the the textureless white paper region has been included by virtue of the height profile of its boundary. Obviously, this could have been done by determining whether the contour motions in reciprocal-polar space lay close to the extracted sinusoid defining the homography, but this does not give any quantitative information about height, which may be necessary if we wanted to allow the robot to drive over obstacles of small height compared to the robot wheel diameter.

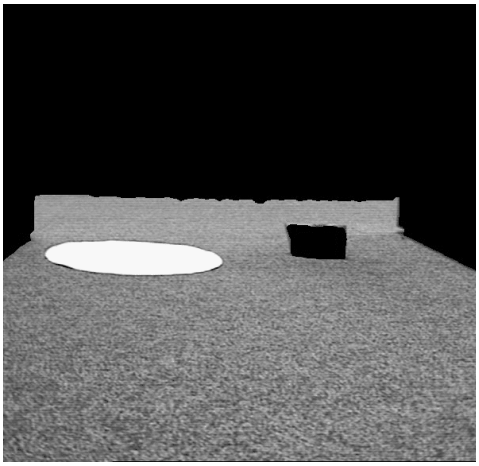

Fig 6. Extracted ground plane.

4 Conclusions

We have described a method which allows a mobile robot's ground plane to be segmented and an affine height landscape constructed by probing the environment with translation manoeuvres. A key point is that all ground plane pixels which have some local variation in intensity/color can be used to contribute to the ground plane homography computation and also we can classify the (transformed) image to be ground plane or non-ground plane at pixel level. The algorithm uses 1D correlations and the robust LS fitting of sinusoids to the resulting shift data to simultaneously recover the ground plane homography and classify pixels. Results have confirmed the validity of both the sinusoid model extraction and affine height measurement procedure.

References

[1] Fischler M. A. and Bolles R. C. Random sample consensus: apardigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, 24:381–395, 1981.

[2] Longuet-Higgins H C. The reconstruction of a plane surface from two perspective projections. *Proc. Royal Society London*, B227:399–410, 1986.

[3] R Hartley and A Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2001.

[4] Faugeras O and Lustman F. Motion and structure from motion in a piecewise planar environment. *Int. Journ. Pattern Recognition and Artificial Intelligence*, 2(3):485–508, 1988.

[5] Tsai R and Huang T. Estimating three-dimensional motion parameters of a rigid planar patch. *IEEE Trans. Acoustics, Speech and Signal Processing*, 29(6):1147–1152, 1981.