EURASIP Journal on Applied Signal Processing 2005:14, 2250–2259

Mobile Robot Visual Navigation Using Multiple Features

Nick Pears
Department of Computer Science, University of York, York Y010 5DD, UK
Email: nep@cs.york.ac.uk

Bojian Liang
Department of Computer Science, University of York, York Y010 5DD, UK
Email: bojian@cs.york.ac.uk

Zezhi Chen
Department of Computer Science, University of York, York Y010 5DD, UK
Email: chen@cs.york.ac.uk

Received 22 December 2003; Revised 29 July 2004

We propose a method to segment the ground plane from a mobile robot's visual field of view and then measure the height of nonground plane features 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, (iii) a novel projective construction for measuring affine height, and (iv) an algorithm that can make use of a variety of visual features and therefore operate in a wide variety of visual environments.

Keywords and phrases: plane segmentation, image rectification, plane and parallax, obstacle detection, mobile robots.

1. INTRODUCTION

1.1. Robust, multifeature, multicue vision systems

In order to operate reliably over extended periods of time (i.e., hours/days/weeks rather than seconds/minutes), computer vision systems must use of all the information in the image stream that is pertinent to the current task. This requires that the system can make this pertinent information explicit by employing a range of feature extractors and visual cues opportunistically, namely, as and when they are available in the image stream and deemed to provide useful constraints to solve task-related problems and resolve any ambiguities. In this way, visual interpretation and decision making can be maximally informed.

We believe that this principle is particularly important in unconstrained environments when the visual environment regularly changes because the disambiguating information content in the image stream is continually changing. Thus if we rely on a single feature/cue combination, such as corners/corner-motion, the application will fail in scenes with few corners, or poorly distributed corners in image space.

We have focused on mobile robot visual navigation as a challenging computer vision problem because the nature of the application suggests that the visual environment is likely to be variable as the robot moves, for example, from room to room. Indeed, we make no assumptions in the work presented here, other than having a reasonably flat floor. This makes the work applicable to mainly indoor mobile robot applications, but also outdoor applications which traverse reasonably flat man-made structures such as pedestrian walkways.

Various visual features (corners, edges, color, texture) and visual cues (e.g., feature motion, parallax) have been employed to facilitate navigational functions with uncalibrated cameras. These include navigation down corridors both by using the focus of expansion of nonvertical scene lines [15] and wide field peripheral flow [3]. Obstacle detection using the projective invariants associated with three horizontal tracked lines and the vanishing lines of planes (ground plane and obstacle planes) has been applied to road scenes successfully [5]. Other approaches to navigation have used time-to-contact from image divergence [14], a combination of central flow divergence and peripheral flow [2], and quantitative planar region detection using point correspondences [12]. Most of these techniques work in some types of scene, but will fail when a particular type of feature/cue combination is not well supported within the image data. We believe the solution is to employ multifeature, multicue approaches.

A point to note is that systems that use this multifeature, multicue philosophy will be highly compute-intensive, due to having a range of processes with a high-bandwidth data input (raw video data). We envisage that the high-bandwidth low-level feature extractors will be implemented as a hardware layer with parallel processing pipelines, either via field-programmable gate arrays (FPGAs) or special purpose DSPs, which output a feature stream. Higher-level layers of software will be able to access the much lower-bandwidth feature stream and select/combine/fuse those parts of the stream, thus providing information or constraints relevant to the particular visual task (in our case robot navigation). Of course, how best to design a framework that allows a computer vision system to understand how to combine features and cues in the context of the current visually-driven task and in the context of the current visual environment is a difficult, open, cross-disciplinary research question. What we aim to do is less ambitious: we aim to develop computer vision applications that are robust, because they are able to make use of multiple features/cues. There is no explicit scene understanding as such, but we need to understand how to build multifeature, multicue algorithms manually before we can develop systems that learn and adapt the way in which they integrate features and cues.

The algorithm that we describe in the following section is able to provide (i) the segmentation of the ground plane or, equivalently, grouping of the ground plane pixels, (ii) a measurement of the height of all other features (corners, edges, or even pixels) above the extracted ground plane. Obviously, such information may be fed directly into a range of obstacle avoidance algorithms (both behavioral and planned path), but may also be used as an input into robot localisation and mapping algorithms.

The algorithm can make use of ground plane and non-ground plane corner features, but if those are scarce (we need at least two corner correspondences), it can make use of any region that has some local intensity variation. If there is no local intensity variation (i.e., if the region is smooth), it can make use of the motion/deformation of the boundary of the smooth region, where boundaries are extracted using a segmentation algorithm based on color-texture properties.

1.2. Outline of a multifeature, multicue robot visual navigation

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 center above the ground plane or alternatively some known height measurement in the scene. 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.

There are several different methods to determine scene structure in the computer vision literature. For example, 3D world structure can be computed from uncalibrated views of a scene given sufficient correspondences in general position and this has already been used to answer specific, metric questions about the scene. The approach by Tomasi and Kanade [8] is known as the factorization method; Triggs [9] extends the factorization method to the projective camera model by using epipolar constraints to calculate depth scale factors; Heyden et al. [11] upgrade the affine approximations to projective results by iterative optimization. Others use the camera-centered approach where the first view is used as the reference camera to determine the projection matrices of other cameras in a projective frame under multiple view geometric constraints. Criminisi et al. [6] proposed methods to make measurements of world planes from their (single) perspective images. Reid and Zisserman [4] give a method for locating 3D position of a soccer ball from monocular image sequence of soccer games.

An important part of our algorithm is that it can use a wide range of features, but, in addition, we present three significant new results: first, by expressing images in a reciprocal-polar $(1/r, \alpha)$ form with the origin on the focus of expansion (FOE),[1] the image motion of a set of coplanar points along the $1/r$ 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 nonground plane pixels, referenced to the height of the camera optical center 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

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

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 [1], Longuet-Higgins [7], and Faugeras and Lustman [10]. 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 $\lambda\mathbf{x}_j = \mathbf{H}x_i$, where $\mathbf{x}$ represents a homogenous image coordinate $(x, y, 1)^T$, $\mathbf{H}$ is a 3×3 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, not necessarily parallel to the plane, a planar homography is often termed a planar homology [16], 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 should lie within a circle of small radius), then translation is assumed and the FOE is computed using random sample consensus (RANSAC [13]) 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 [16]. Otherwise, the FOE is at a distance $d = q/\sqrt{a_v^2 + b_v^2}$ from the vanishing line. To simplify notation, we now drop the "prime" notation from (3) and assume that (x, y) are image measurements made relative to the FOE. Thus, we have

$$\begin{aligned} (-ka_vx_1 - kb_vy_1 + 1 - kq)x_2 &= x_1, \\ (-ka_vx_1 - kb_vy_1 + 1 - kq)y_2 &= y_1. \end{aligned} \tag{4}$$

Squaring both sides and adding,

$$\left(-k(a_vx_1 + b_vy_1) + (1 - kq)\right)^2(x_2^2 + y_2^2) = (x_1^2 + y_1^2). \tag{5}$$

If we define r_i as the Euclidean distance between an image point and the FOE in frame i, then, taking square roots of (5),

$$\begin{aligned} \left(-k(a_vx_1 + b_vy_1) + (1 - kq)\right)r_2 &= r_1, \\ r_2^{-1} &= -k(a_vx_1 + b_vy_1)r_1^{-1} + (1 - kq)r_1^{-1}, \\ r_2^{-1} &= -k(a_v\cos\alpha + b_v\sin\alpha) + (1 - kq)r_1^{-1}, \end{aligned} \tag{6}$$

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

$$a_v = -\sqrt{(a_v^2 + b_v^2)}\sin\alpha_v, \quad b_v = \sqrt{(a_v^2 + b_v^2)}\cos\alpha_v. \tag{7}$$

Hence

$$r_2^{-1} = k_p(-\sin\alpha_v\cos\alpha + \cos\alpha_v\sin\alpha) + k_qr_1^{-1}, \tag{8}$$

where

$$k_p = -k\sqrt{(a_v^2 + b_v^2)}, \quad k_q = 1 - kq. \tag{9}$$

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

$$s(\alpha) = \rho_2 - k_q\rho_1 = k_p \sin(\alpha - \alpha_v), \tag{10}$$

where we define $\rho = 1/r$. Equation (10) 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 (10) 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

$$s_i = k_p \sin(\alpha_i - \alpha_v), \quad s_j = k_p \sin(\alpha_j - \alpha_v), \tag{11}$$

hence

$$\begin{aligned} \frac{s_i}{s_j} &= \frac{\sin\alpha_i \cos\alpha_v - \cos\alpha_i \sin\alpha_v}{\sin\alpha_j \cos\alpha_v - \cos\alpha_j \sin\alpha_v}, \\ \frac{s_i}{s_j} &= \frac{\sin\alpha_i - \cos\alpha_i \tan\alpha_v}{\sin\alpha_j - \cos\alpha_j \tan\alpha_v}. \end{aligned} \tag{12}$$

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

$$\tan\alpha_v = \frac{s_j \sin\alpha_i - s_i \sin\alpha_j}{s_j \cos\alpha_i - s_i \cos\alpha_j}. \tag{13}$$

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 (11). 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 [13] 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 an 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, coplanar pixels may be grouped without explicit construction of a homography matrix. Now, let

$$s_i = k_p \sin(\alpha_i - \alpha_v) = m \sin\alpha_i - n \cos\alpha_i, \tag{14}$$

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

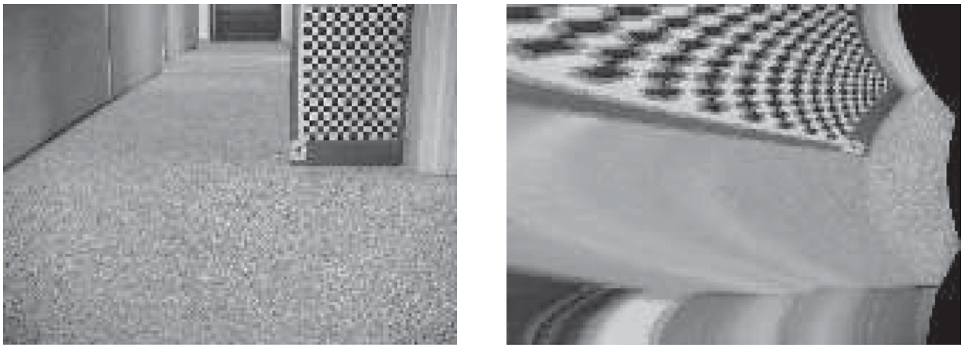
$$\begin{bmatrix} -\sin\alpha_i & \cos\alpha_i & s_i \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} m \\ n \\ 1 \end{bmatrix} = \mathbf{0}. \tag{15}$$

We use singular value decomposition (SVD) to solve for $\lambda[m, n, 1]^T$ and normalise the solution to obtain the parameters, (m, n). From (3), we substitute $n = -ka_v$, $m = -kb_v$, and $k_q = 1 - kq$, to recover the homography in an FOE centered frame. The parameters defining $s(\alpha)$ can be computed as $k_p = \sqrt{(m^2 + n^2)}$, $\alpha_v = \tan^{-1}(-n/m)$ 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 (this may be a simple linear interpolation or based on cubic splines) to RP (ρ, α) form. Figure 1a shows one of the original images which has its FOE sixteen pixels above the top center edge of its image. Its RP transform is in Figure 1b and shows the angular (α) axis and lines of constant (reciprocal) radius in the vertical direction. The axis representing (scaled) reciprocal distance from the FOE (ρ) and reciprocal radial lines (constant angle) are in the horizontal direction. For a more intuitive viewing, the horizontal rendering of the RP image is such that r increases from left to right, thus ρ increases from right to left. In this rendering of the RP image, the FOE is out at infinity to the left of the image. In order to usefully constrain the size of the RP image, any pixels less than 64 pixels from the FOE are not included in the RP plot. In any case, the motion near the FOE is so small that it is not possible to make accurate height measurements in this image region. In terms of implementation details, we used VGA-sized images (640 × 480) as the raw input images and we generated RP images such that there was no information loss due to pixel compression.

For each pixel in the original image (2nd of pair), we find its position in RP-space, and find the maximum correlation



(a)

(b)

Figure 1: (a) shows an original image $I(x, y)$ and (b) shows the corresponding RP image $I(\rho, \alpha)$.

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 $s(\alpha)$. A correlation window size of 64 pixels was used in the RP image, with a search window of 150 pixels. Due to the high-frequency repetitive structure of the carpet microtexture shown in Figure 1a, often many local maxima are generated in the correlation process. However, due to low-frequency components of spatial frequency, the global maximum, in most cases, corresponds to the correct image motion that we wish to recover. Even if this is not the case, it is possible to include pixels with significant local maxima, if these strong local maxima are consistent with the extracted planar motion, that is, they lie sufficiently close to the extracted sinusoid model.

The plot of $s(\alpha)$ against α is shown in Figure 2a 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 used to compute the ground plane homography. The inliers of the recovered sinusoid are then plotted in 3D in Figure 2b, 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, as expected.

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 center 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 $\mathbf{b} = \mathbf{Ha}$. Referring to Figure 3, the height of the feature point relative to the height of the camera optical center, the affine height, is

$$h_a = \frac{h}{h_c} = 1 - \frac{D}{h_c}. \tag{16}$$

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

$$\frac{D}{h_c} = \frac{d(OC)}{d(OC')} = \frac{d(AC)}{d(A'C')}. \tag{17}$$

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

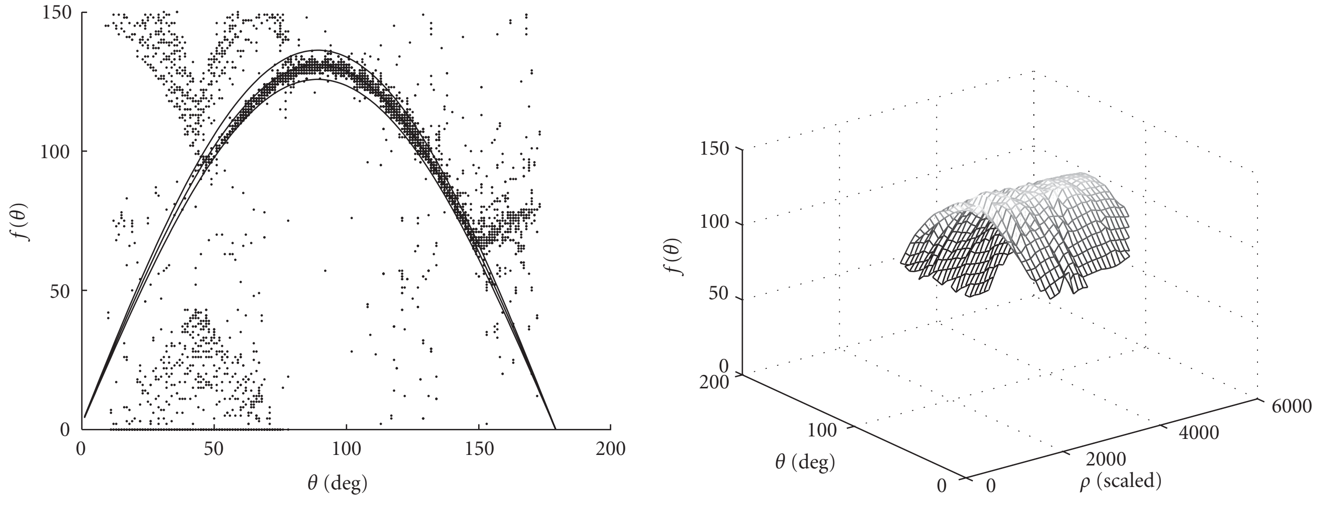
$$h_a = 1 - \frac{d(A'B')}{d(A'C')}. \tag{18}$$

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

$$\frac{d(A'B')}{d(A', C')} = \frac{d(a, b)d(c, x_f)}{d(a, c)d(b, x_f)}. \tag{19}$$

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

$$h_a = 1 - \frac{d(a, b)d(c, x_f)}{d(a, c)d(b, x_f)}. \tag{20}$$



(a)

(b)

FIGURE 2: (a) All image motion in RP-space. (b) Ground plane image motion in RP-space with ρ plotted explicitly.

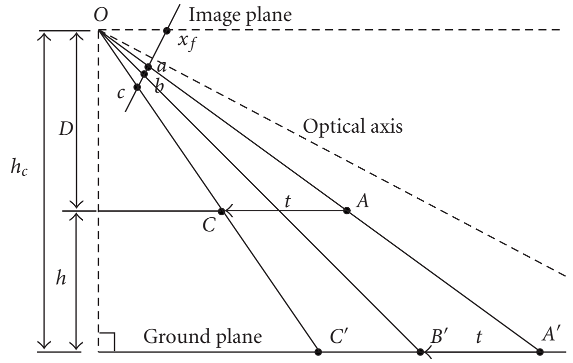


FIGURE 3: Measuring the height of point A.

In general, we find that

$$h_a = 1 + \mu \frac{d(a,b)d(c,x_f)}{d(a,c)d(b,x_f)}, \tag{21}$$

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.

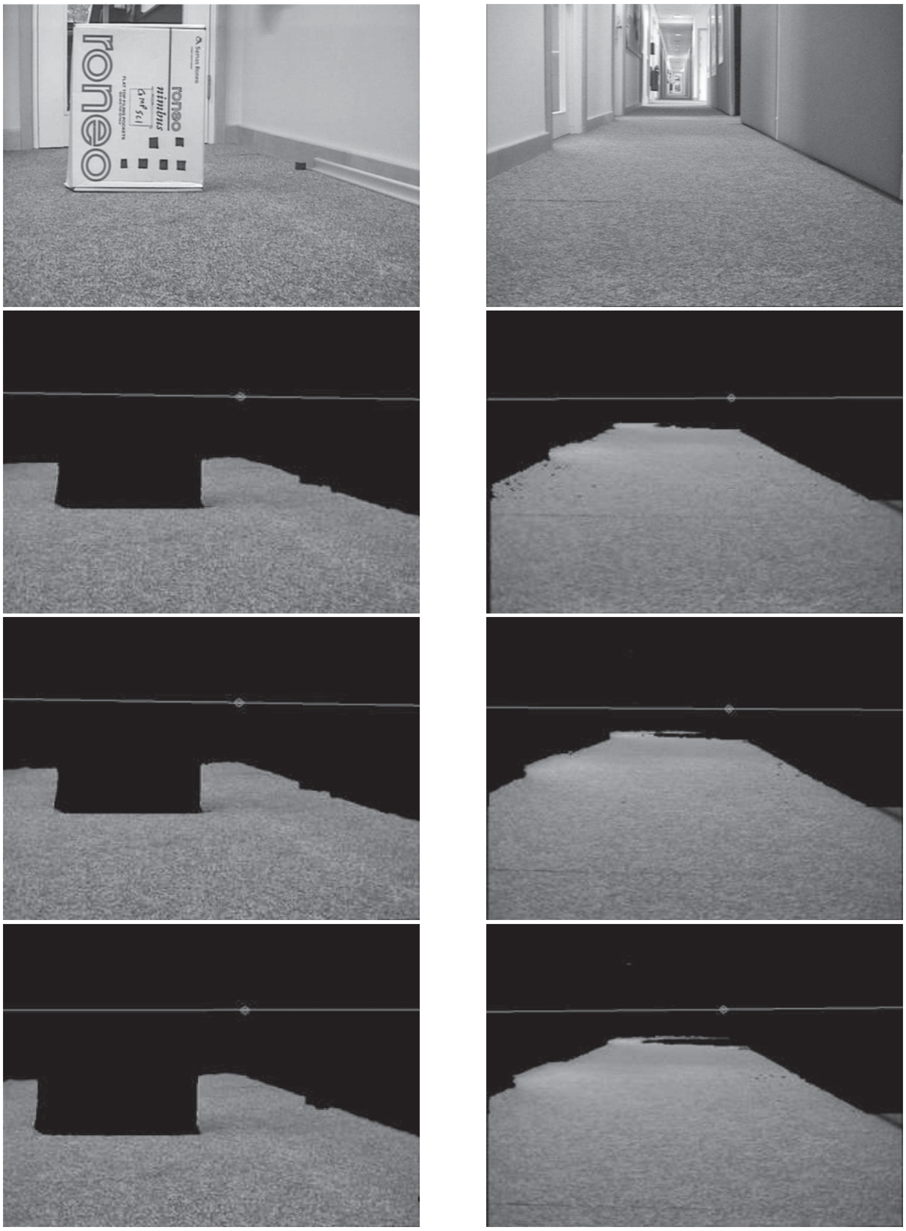
4. EXPERIMENTAL RESULTS

4.1. *Ground plane segmentation by ground plane pixel grouping*

In the two experiments shown in Figure 4, the ground plane was segmented on a pixel-by-pixel basis, as the robot translated forwards. Those pixels that fit the ground plane sinusoid in RP image space are retained and plotted as ground plane pixels, all other pixels are removed from the image (rendered as zero intensity, or black). Comparing the segmentations with their original images, it can be seen that the pixel-by-pixel grouping gives an accurate ground plane segmentation. For example, the small, black door stop on the center-right of the first sequence is clearly and correctly excluded. Also, the small piece of carpet at the foot of the door in the center-left of the second sequence is clearly and correctly included. Note also that the extracted FOE is shown as a small circle in both sequences and the vanishing line of the ground plane, whose orientation is extracted as the phase of the RP sinusoid is also shown as a near horizontal line (near zero sinusoid phase).

4.2. *Height measurement experiments*

Two experiments (one indoor, one outdoor) are presented to validate the projective construction used to measure height. In both experiments, $q = 0$ and equivalently $k_q = 1$ was assumed, since two VGA frames (640×480 resolution) were captured with the translation direction parallel to the ground plane. In both cases the height A-B was used as the reference height to correctly scale affine (relative) height measurements to Euclidean (absolute) height measurements. For the indoor experiment, we computed the orientation of the vanishing line as $\tan \alpha_v = 0$, and for the outdoor experiment



(a)

(b)

Figure 4: Image sequences showing ground plane segmentation. (a) Image sequence 1. (b) Image sequence 2.

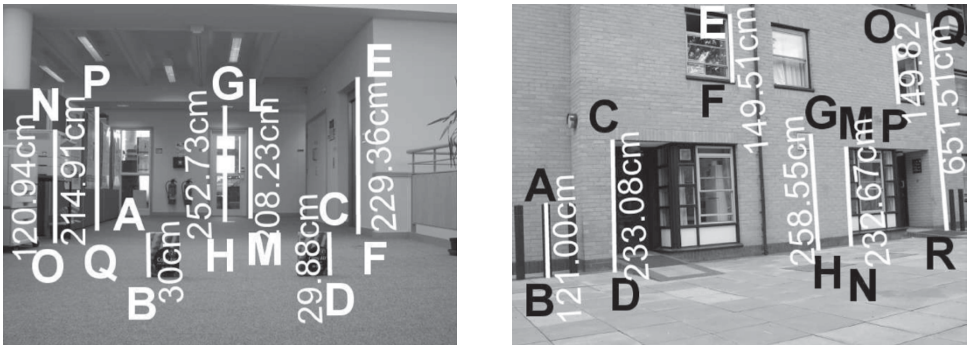
we computed $\tan \alpha_v = -0.03$. Results are shown in Figure 5 and in Table 1, where "TM" are the manual (tape measure) measurements and "VM" are the results from our automatic height measurement method. We find a mean absolute error of 6.9 mm and mean relative error of 0.35%. If we remove the two rather inaccurate measurements (a)EF and (a)PQ, the remaining measurements have a mean absolute error of 1.5 mm and a 0.1% mean relative error.

Table 1: Height measurement results in centimetres.

Segment	TM	VM	Segment	TM	VM
(a)CD	30.0	29.88	(b)CD	233.1	233.08
(a)EF	227.7	229.36	(b)EF	149.8	149.51
(a)LM	208.4	208.23	(b)GH	258.7	258.55
(a)GH	252.5	252.73	(b)MN	233.1	232.67
(a)NO	121.1	120.94	(b)OP	149.8	149.82
(a)PQ	210.3	214.91	(b)QR	None	651.51

4.3. Using height profiles of smooth regions to segment the ground plane

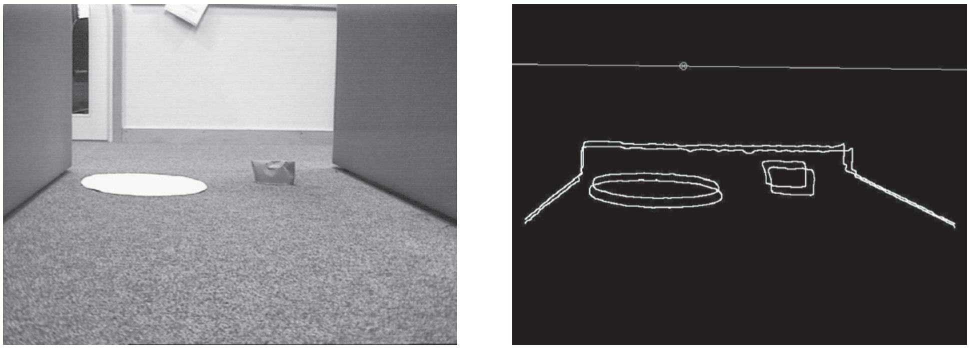
The final experiment presented in this paper uses both the sinusoid fitting process (to simultaneously recover the textured regions of the ground plane and the associated ground plane homography) *and* the affine height measurement method to determine whether the contours of *nontextured* regions



(a)

(b)

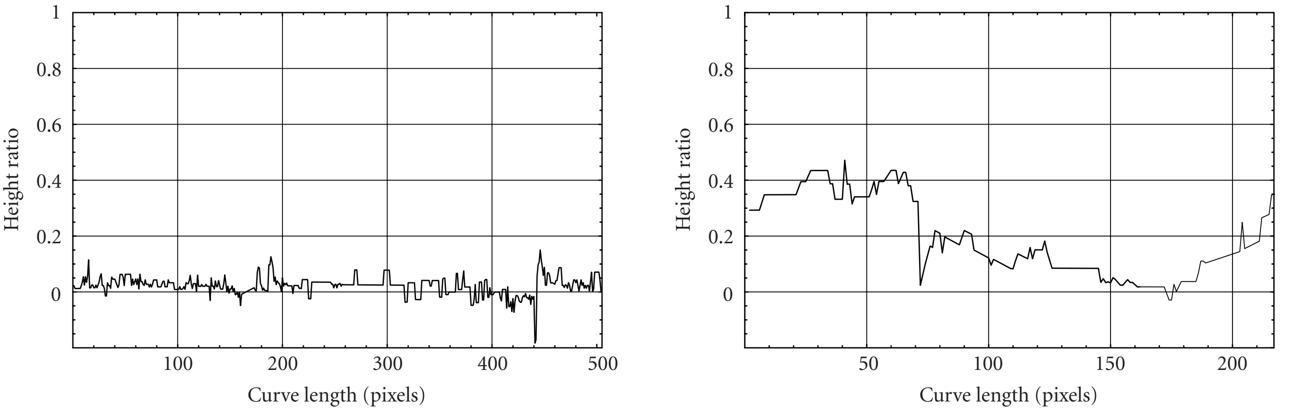
Figure 5: (a) Indoor height measurement experiment, and (b) outdoor height measurement experiment.



(a)

(b)

Figure 6: (a) Raw image. (b) Boundaries to be matched using recovered FOE.



(a)

(b)

Figure 7: (a) Height profile of coplanar white paper. (b) Height profile of small box.

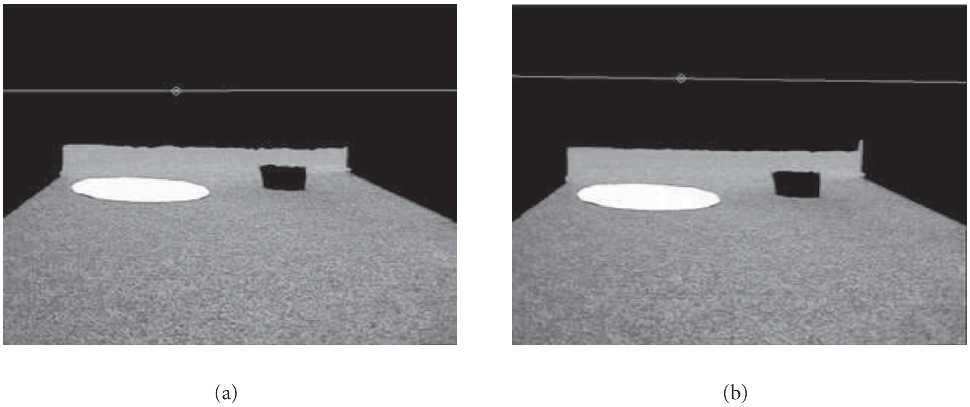


Figure 8: Two frames of the extracted ground plane. (a) Height profile of white paper. (b) Height profile of small box.

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 color-texture. Textureless regions cannot be classified as ground plane or nonground 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 by casting rays from the FOE recovered in the homography estimation process.

Figure 6a 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 not. Figure 6b shows the extracted boundaries and the FOE 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 Figure 7a. Otherwise it is classified as an obstacle, as in Figure 7b. The final image in Figure 8, shows two frames of the extracted ground region where the textured carpet has been classified on a pixel-by-pixel basis, and 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.

5. 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 nonground 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. The approach may be used for a range of feature types including corners, edges, region boundaries, and even raw pixels if they have some local color-intensity variation.

REFERENCES

[1] R. Tsai and T. Huang, "Estimating three-dimensional motion parameters of a rigid planar patch," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 6, pp. 1147–1152, 1981.

[2] D. Coombs, M. Herman, T.-H. Hong, and M. Nashman, "Real-time obstacle avoidance using central flow divergence, and peripheral flow," *IEEE Trans. Robot. Automat.*, vol. 14, no. 1, pp. 49–59, 1998.

[3] J. Santos-Victor, G. Sandini, F. Curotto, and S. Garibaldi, "Divergent stereo in autonomous navigation: from bees to robots," *International Journal of Computer Vision*, vol. 14, no. 2, pp. 159–177, 1995.

[4] I. D. Reid and A. Zisserman, "Goal-directed video metrology," in *Proc. European Conference on Computer Vision (ECCV '96)*, R. Cipolla and B. Buxton, Eds., vol. 2, pp. 647–658, Cambridge, UK, April 1996.

[5] R. Okada, Y. Taniguchi, K. Furukawa, and K. Onoguchi, "Obstacle detection using projective invariant and vanishing lines," in *Proc. IEEE 9th International Conference on Computer Vision (ICCV '03)*, pp. 330–337, Nice, France, October 2003.

[6] A. Criminisi, I. D. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

[7] H. C. Longuet-Higgins, "The reconstruction of a plane surface from two perspective projections," *Proceedings of Royal Society of London B*, vol. 227, pp. 399–410, 1986.

[8] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorisation approach," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[9] B. Triggs, "Factorization methods for projective structure and motion," in *Proc. IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition (CVPR '96), pp. 845–851, San Francisco, Calif, USA, June 1996.

[10] O. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, no. 3, pp. 485–508, 1988.

[11] A. Heyden, R. Berthilsson, and G. Sparr, "An iterative factorization method for projective structure and motion from image sequences," *Image and Vision Computing*, vol. 17, no. 13, pp. 981–991, 1999.

[12] D. Sinclair and A. Blake, "Quantitative planar region detection," *International Journal of Computer Vision*, vol. 18, no. 1, pp. 77–91, 1996.

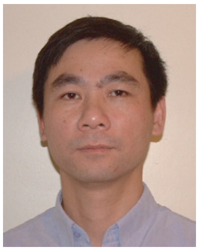
[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: apardigm for model fitting with application to image analysis and automated cartography," *Communications of the Associa-tion for Computing Machinery*, vol. 24, pp. 381–395, 1981.

[14] R. Cipolla and A. Blake, "Surface orientation and time to contact from image divergence and deformation," in *Proc. 2nd European Conference on Computer Vision (ECCV '92)*, pp. 187–202, Santa Margherita Ligure, Italy, May 1992.

[15] J. J. Guerrero and C. Sagüés, "Navigation from uncalibrated monocular vision," in *Proc. 3rd IFAC Symposium on Intelligent Autonomous Vehicles*, pp. 210–215, Madrid, Spain, March 1998.

[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, New York, NY, USA, 2001.

Nick Pears was awarded both a B.S. degree in engineering science and a Ph.D. degree in robotics in 1990 by Durham University, UK. He then worked in the Robotics Research Group, Oxford University, and in 1994 he was elected a Fellow of Girton College, Cambridge University. In 1998, he joined the Computer Science Department, University of York. His current research interests include visual navigation, face recognition, visual human-computer interaction, and visual metrology.



Bojian Liang was awarded a B.S. degree in communication engineering from Northern Jiaotong University, China, and a Ph.D. degree in computer science from Heriot-Watt University, Edinburgh. Recent research interests include object recognition, robot navigation, and the application of polarization analysis to computer vision problems. He is currently working on the DAME project at the University of York where his specialization is pattern matching and time series signal analysis.

Zezhi Chen was awarded an M.S. degree in computer aided geometric design by Northwest University, Xi'an, China. He received a Ph.D. degree in communication and information engineering from Xidian University, Xi'an, China, in 2002. He now works in the Department of Computer Science, University of York, UK. His current main research interests are 3D reconstruction, applications of computer vision, and computer graphics.