Plane Segmentation from Two Views in Reciprocal-Polar Image Space

Zezhi Chen, Nick E. Pears, Bojian Liang, and John McDermid

Department of Computer Science, University of York, YO10 5DD, UK
{chen,nep,bojian,jam}@cs.york.ac.uk

Abstract. We present a new method of segmenting planar regions when an uncalibrated camera undergoes (near) pure translation. We show that, for pure translation parallel to a plane, the relation between the two views, when expressed in a reciprocal-polar $(\frac{1}{r}, \theta)$ space, is a pure shift in the $(\frac{1}{r})$ dimension for a given value of θ. Furthermore, we show that the magnitude of these shifts follows a sinusoidal form along the θ direction over a maximum of π radians. This allows planar image motion to be accurately detected and recovered by 1D correlation. Simultaneous planar pixel grouping and recovery of the plane homography, thus amounts to a robustly fitting a sinusoid to shifts of maximum correlation in reciprocal-polar space. The phase of the recovered sinusoid corresponds to the orientation of the vanishing line of the plane and the amplitude is related to the magnitude of the camera translation.

1 Introduction

In this work[1], a new method of segmenting planar regions when a camera undergoes pure translation is presented. It has been demonstrated as a practical approach to monocular mobile robot obstacle detection or, equivalently, mobile robot ground plane segmentation. The essence of the paper is the reciprocal-polar image (and the use of sinusoid models to segment planes within that space), which is new, although related ideas have been presented in the literature. For example, Pollefeys et al [3] suggested a polar rectification (r, θ) to aid stereo matching. Wolberg and Zokai [4] have used the well-known log-polar transformation to aid affine motion recovery. Both of these allow more general motion than our transformation, but do not give the main benefit of the reciprocal-polar transformation, which allows correlation based matching over large camera motions. The use of intensity correlation means that there can be low density (or even zero) corner features on the plane of interest (eg ground plane). Thus our method gives good performance compared to other plane segmentation schemes particularly when the camera motion is large and the density of corner features on the plane of interest is low. We summarise the advantages as follows:

[1] The authors acknowledge the support of the UK DTI Aeronautics Research Programme and the EPSRC

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 638–646, 2004.

1. It renders image motion, for a given radial direction on a planar surface, into a pure shift. This allows image motion to be recovered by direct correlation methods (in addition to the usual feature correspondences), which fail in the original image space for large image motions, due to perspective distortion. In reciprocal-polar space, for a given epipolar line (i.e radial direction relative to the focus of expansion), distant points on a plane (i.e. points close to the focus of expansion) move the same amount as points on the same plane, which are close to the camera.
2. The amount of radial image motion expressed in reciprocal-polar space, follows a sinusoidal variation, where the phase of the sinusoid is the orientation of the vanishing line of the plane (plane horizon) and the magnitude of the sinusoid is dependent on the magnitude of the translation, camera intrinsic parameters and the proximity of the plane. This allows planar grouping down to pixel level, using a simple sinusoid model, given that there is some local texture around the pixel. In terms of (robust) least squares extraction of models (here, homographies), you get better solutions if you can use more data associated to that model. This is why our method can improve on methods based purely on feature matching approaches, particularly in scenes where the features on the plane of interest (ground plane) are sparse.
3. The method works even in the absence of feature correspondences on the ground plane, although feature correspondences are not precluded from being classified as ground plane or obstacle using the sinusoid model. Indeed, in the absence of local texture, we cannot segment on a pixel basis, and contours of smooth regions can be matched (using the focus of expansion, or epipole) and classified using the sinusoid model.
4. Compared to other approaches in mobile robot visual navigation, it is simple to distinguish the ground plane from other planes in the scene, using some a-priori knowledge of the camera roll orientation with respect to the ground plane. Usually the roll angle is very close to zero, so the vanishing line and phase of the sinusoid are close to zero. Planes at other orientations relative to the ground plane have different phase sinusoids. Planes parallel to the ground plane (eg tops of obstacles) and below the camera have the same phase but all have a larger amplitude sinusoid than the ground plane. Planes above the camera and parallel to the ground plane (eg the ceiling, if visible) have a phase shift of π relative to the ground plane's sinusoid model.

2 The Relation Between Planar Homography (H) and Fundamental Matrix (F) Under Pure Translation

In pure translation of the camera, one may consider an equivalent situation in which the camera is stationary, and the world undergoes a translation $-\mathbf{t}$. In this situation, points in 3D space move on straight lines parallel to $\mathbf{t}$, and the imaged intersection of this parallel lines is the focus of expansion (FOE) $\mathbf{v}$ in the direction of $\mathbf{t}$. It is evident that $\mathbf{v}$ is also the epipole $\mathbf{e}$ and $\mathbf{e}'$ for both views, and the imaged parallel lines are epipolar lines.

$$\mathbf{F} = [\mathbf{e}']_\times = [\mathbf{e}]_\times = [\mathbf{v}]_\times . \tag{1}$$

where $[\bullet]_\times$ is a skew-symmetric matrix corresponding to the vector.

Let $\mathbf{X}_i$ be a set of points which are coplanar in 3-D Euclidean space. The images of $\mathbf{X}_i$ from two view points are related by a plane to plane projectivity or homography $\mathbf{H}$, such that, $\lambda \mathbf{x}_{i2} = \mathbf{H}\mathbf{x}_{i1}$. where λ is a scalar, $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are homogenous image coordinates of the images of point $\mathbf{X}_i$, $\mathbf{H}$ is a 3×3 matrix representing the homography. Note that two corresponding point pairs fully determine the H-matrix under pure translation parallel to the plane. Suppose the cameras are calibrated with the world origin at the first camera and the intrinsic parameters constant.

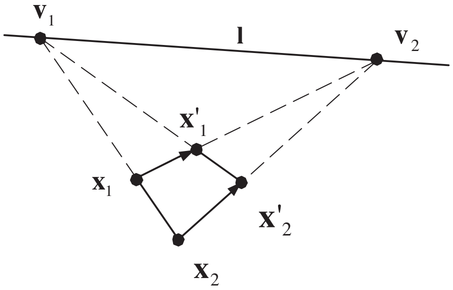


Fig. 1. Two corresponding point pairs fully define the FOE and vanishing line

$$\mathbf{P} = \mathbf{K}[\mathbf{I}|\mathbf{0}] \qquad\qquad \mathbf{P}' = \mathbf{K}[\mathbf{R}|\mathbf{T}]. \tag{2}$$

and the world plane π_E has coordinates $\pi_E = (\mathbf{n}^T, d)^T$ so that $\mathbf{H} = \mathbf{K}(\mathbf{R} - \lambda \mathbf{t}\mathbf{n}^T)\mathbf{K}^{-1}$. Where $\lambda = \frac{1}{d}$, For a pure translation, $\mathbf{R}$=$\mathbf{I}$, and so $\mathbf{H}$ has the form $\mathbf{H} = \mathbf{I} - \lambda(\mathbf{K}\mathbf{t})(\mathbf{K}^{-T}\mathbf{n})^T$. We note that $\mathbf{K}\mathbf{t}$ is the FOE $\mathbf{v} = \eta(x_f \quad y_f \quad 1)^T$, and $\mathbf{K}^{-T}\mathbf{n}$ is the vanishing line $\mathbf{l} = v(a_v \quad b_v \quad 1)^T$ corresponding to plane π_E . Thus we have

$$\mathbf{H} = \mathbf{I} - k\mathbf{v}\mathbf{l}^T. \tag{3}$$

Where k is a constant scalar. Because two corresponding point pairs fully define the FOE and vanishing line, so H-matrix can, in theory, also be fully determined by the two corresponding matches (Fig. 1). However, in our robust method, we recover the FOE using the algorithm given in the following section. We then determine the vanishing line orientation using the phase of a sinusoid fitted to feature matches and/or correlation maxima in reciprocal-polar space, using RANSAC and least-squares (SVD).

3 Recovering the Focus of Expansion

We can detect (near) pure translation by intersecting all lines defined by *all* corner correspondences from the image pair and if a large percentage of intersections lie in a small area (such as 90% of intersections should lie within a 50 pixels radius), then pure translation is assumed and the FOE is computed. The question now is: how can we calculate the FOE with high accuracy and high stability? We have developed a robust method which uses a corner matching procedure and then uses random sample consensus (RANSAC) and standard linear methods (SVD) to solve the overdetermined system of linear equations $\mathbf{x}_i \times \mathbf{x}'_i) \bullet \mathbf{v} = 0$, where $\mathbf{x}_i \longleftrightarrow \mathbf{x}'_i$ $(i = 1, 2, \cdots, n)$ is any pair of matching points in two images and $\mathbf{v}$ is the required FOE. After this linear process an iterative procedure is used to minimise the epipolar distance, $f(\mathbf{v})$, given by:

$$f(\mathbf{v}) = \left(\frac{1}{\sqrt{(\mathbf{F}\mathbf{x}_i)_1^2 + (\mathbf{F}\mathbf{x}_i)_2^2}} + \frac{1}{\sqrt{(\mathbf{F}^T\mathbf{x}'_i)_1^2 + (\mathbf{F}^T\mathbf{x}'_i)_2^2}} \right) \left| {\mathbf{x}'_i}^T \mathbf{F}\mathbf{x}_i \right|$$

where $\mathbf{F}$ is the fundamental matrix.

4 Plane Segmentation and Vanishing Line Recovery

Once the FOE has been computed, we shift image coordinates so that each image is centred on the FOE, using a centering tranlslation matrix, $\mathbf{T}_c$. The FOE is then at homogenous coordinates $\mathbf{v}' = (0,0,1)^T$ and vanishing line becomes $\mathbf{l}' = \mathbf{T}_c^{-T}\mathbf{l} = (a_v, b_v, \mathbf{v}^T\mathbf{l})^T$. The homography relating points in FOE centred coordinates is

$$\mathbf{H}' = \mathbf{I} - k\mathbf{v}'\mathbf{l}'^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -ka_v & -kb_v & (1 - k\mathbf{v}^T\mathbf{l}) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ s & \mu & q \end{bmatrix}. \tag{4}$$

where $q = 1 - k\mathbf{v}^T\mathbf{l}$, $s = -ka_v$ and $\mu = -kb_v$. If the robots translation direction is parallel to the ground, $q = 1$, since the FOE lies on the vanishing line for this special motion. In this specialisation, original H-matrix has four *dof*, and is sometimes termed an elation[2]. Under the new homography, we have

$$(sx'_1 + \mu y'_1 + q)^2 (x'^2_2 + y'^2_2) = (x'^2_1 + y'^2_1). \tag{5}$$

If we define $\rho = \frac{1}{r}$, where r is the Euclidean distance between an image point and the FOE in a frame, then taking square roots of Eq.(5)

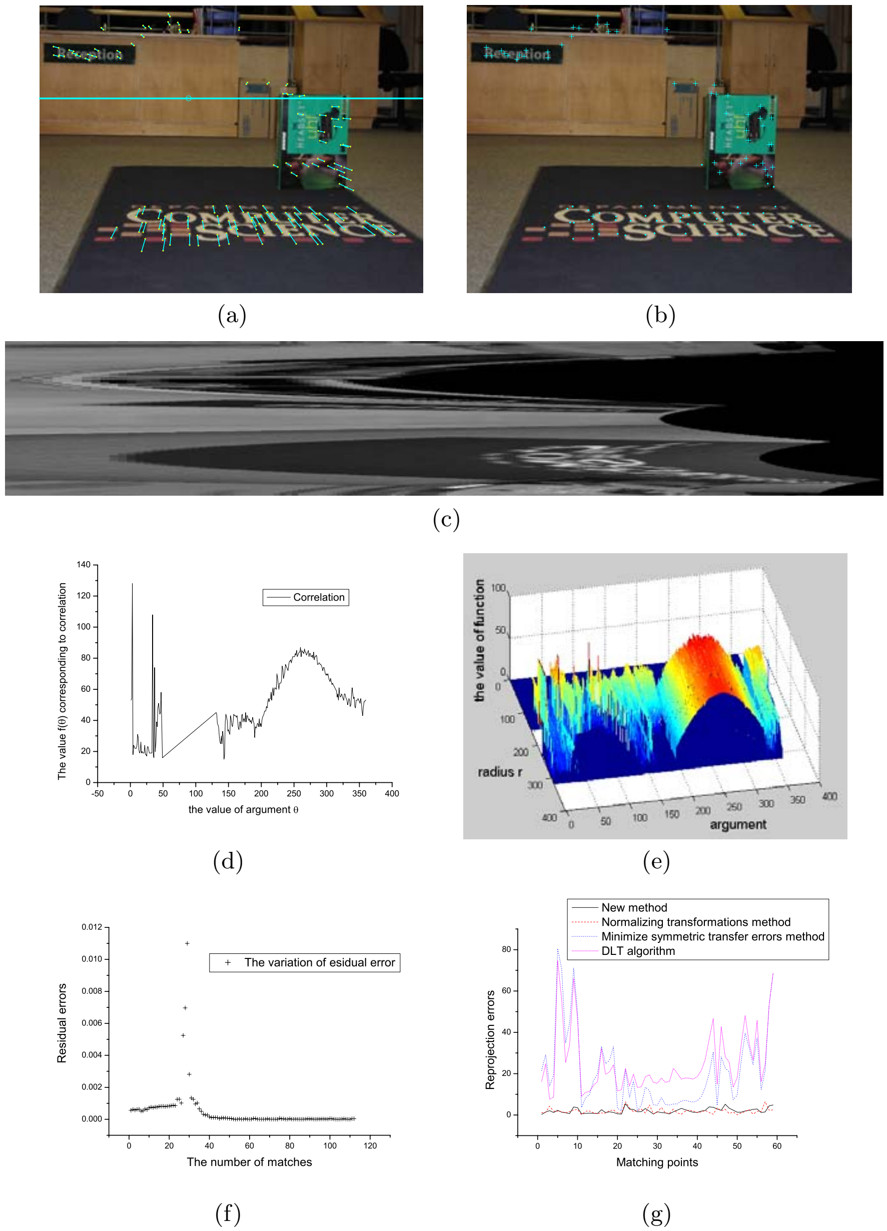
$$f(\theta) = \rho_2 - q\rho_1 = s\frac{{x'}_1}{r_1} + \mu\frac{{y'}_1}{r_1} = k_{s\mu}\sin(\theta + \alpha). \tag{6}$$

where θ is the angular position of a pixel in a frame centred on the FOE.

$$k_{s\mu} = \sqrt{s^2 + \mu^2}, \sin\alpha = \frac{s}{k_{s\mu}}, \cos\alpha = \frac{\mu}{k_{s\mu}}, \tan\alpha = \frac{a_v}{b_v}. \tag{7}$$

Eq.(6) indicates that we need to find three constants $(q, k_{s\mu}, \alpha)$ in order to recover the homography and that the computation should be implemented in (ρ, θ) image space (note that a planar homology has five *dof*, but two have been recovered in the FOE computation). We call this reciprocal-polar image space. Thus, after computing the FOE, a cubic-spline based interpolation procedure is used to generate a reciprocal-polar image for each image in the image pair.

For each pixel in image 1, its position in reciprocal-polar image space is computed, and a 1D window is created around this position along the $q\rho$ dimension ($q = 1$ for translation parallel to the ground plane, for other translations, q is easily computed). We then correlate this window along the ρ in reciprocal-polar image 2, at the same value of θ. This correlation process is possible because of



(a) (b)

(c)

(d) (e)

(f) (g)

Fig. 2. (a) one of the original images with their matching points (•), feature tracks, FOE and vanishing line. (b) ground plane points (•) and obstacle points (+). (c) reciprocal-polar image corresponding to (a). (d) The value of $f(\theta)$. (e) 3D sinusoid. (f) The variation of residual errors (the number of coplanar points is between 54 and 112). (g) Reprojection errors of several alternative methods.

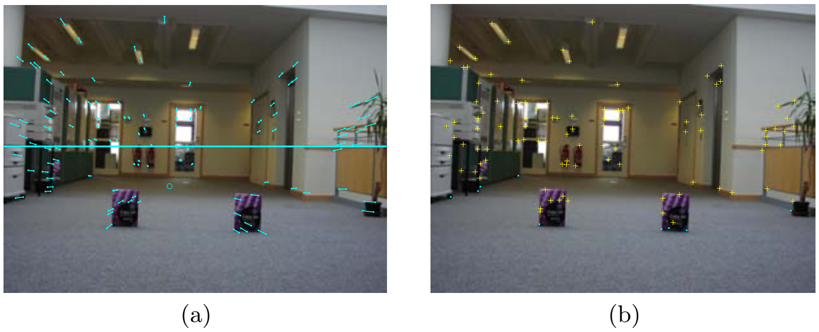


Fig. 3. (a) Correspondences (b) The segmented ground plane points (•) and obstacle points (+)

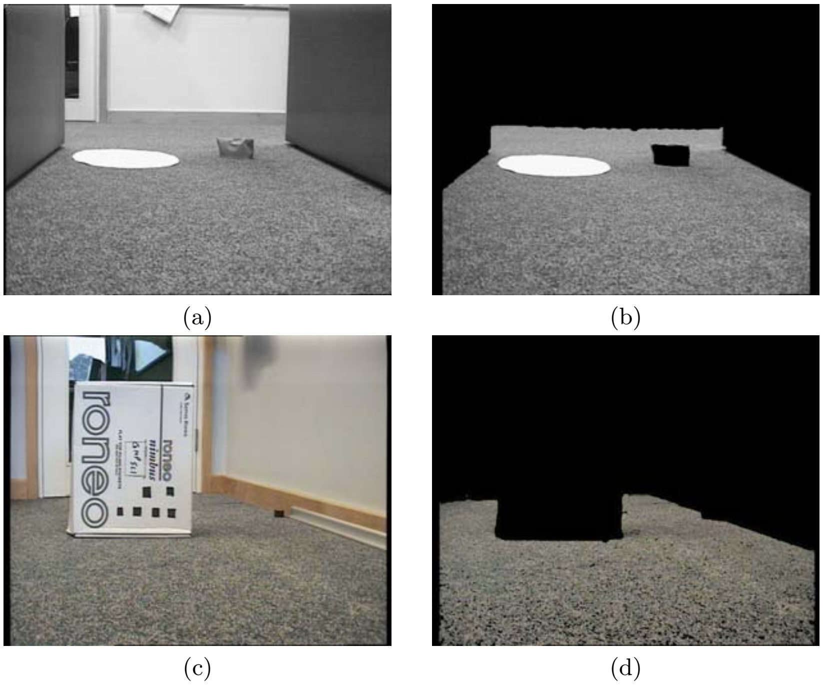


Fig. 4. (a) Original image 1 (b) The segmented ground plane region (c) Original image 2 (b) The segmented ground plane region

the 'pure-shift' relation between ρ_2 and $q\rho_1$, expressed in Eq.(6) and the position of the maximum value of the correlation is related as a value of $f_i(\theta)$.

Eq.(6) indicates that correlation maxima and feature correlations in reciprocal-polar space, which are associated with a planar surface, lie on a sinusoid and the constants $(k_{s\mu}, \alpha)$ may be recovered by fitting a sinusoid to the data for $f(\theta)$. Suppose that we have two value of $f(\theta)$, $f_{i,j}$ measured at two angles, $\theta_{i,j}$, so that

$$f_i = k_{s\mu}\sin(\theta_i + \alpha), f_j = k_{s\mu}\sin(\theta_j + \alpha). \tag{8}$$

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

$$\tan\alpha = \frac{f_j \sin\theta_i - f_i \sin\theta_j}{f_i \cos\theta_j - f_j \cos\theta_i}. \tag{9}$$

Thus, in theory, a pair of f 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. Then, given the phase angle, α, corresponding to the orientation of the vanishing line, we can compute $k_{s\mu}$ from the Eq.(8).

In order to robustly and accurately estimate the vanishing line orientation from many correlation maxima and feature correspondences in reciprocal-polar space, many of which will not be associated with the ground plane, a random sample consensus (RANSAC) [1] and least-squares process is used. Pairs of (θ, f) values are selected across random values of ρ and θ, with the constraint that there must be a minimum angular separation between the two θ values. The magnitude and phase of a sinusoid that passes through these two (θ, f) coordinates is computed and the number of inliers stored. An inlier is defined as some image motion (correlation maxima or feature correspondence) that is within threshold of the putative sinusoid. Thus $r < \mathit{threshold}$, where r is the residual error: $r = |(\rho_{P'} - q\rho_P) - k_{s\mu}\sin(\theta + \alpha)|$ and $\mathbf{P}$ and $\mathbf{P}'$ are correspondences. The sinusoid with the maximum number of inliers is used to initialise an iterative procedure where a standard linear least-squares estimate of the sinusoid parameters and the associated set of inliers are computed until the inlier distribution 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, although this is easily recovered in the FOE centered frame from the FOE and the two parameters (amplitude and phase) of the sinusoid. Note that the FOE and phase of the sinusoid give the vanishing line, and the amplitude of the sinusoid gives the scale of the translation, related to k, hence we have all the parameters that we need in equation 3. The homography in the original frame, if required, can be computed as $\mathbf{H} = \mathbf{T}_c^{-1}\mathbf{H}'\mathbf{T}_c$.

5 Experimental Results and Applications

5.1 Results Using Corner Correspondences

In the first experiment, the camera was moved parallel to the ground plane. Fig. 2 (a) shows one of the original images with their matching points (●), feature tracks, FOE and vanishing line. After the FOE was obtained, the images were then converted to reciprocal-polar (ρ, θ) form, as shown in (c).(d) shows us the one of the sinusoidal forms of $f(\theta)$ for a fixed ρ (and hence fixed r). The partial sinusoidal curve ($\theta \in [193.25, 315.00]$) is clearly shown and represents the motion of the ground plane in reciprocal-polar space. (The phase is shown close to 180 degrees rather than 0 degrees because the direction of y in the image is directed upwards from the FOE rather than downwards). The 3D sinusoidal form of reciprocal-polar ground plane motion, which is obtained when ρ (and hence r) is varied, is shown in (e). In this set of experiments, the number of coplanar matching points varied from 54 to 112. The mean of residual errors of ground plane points is 1.558E-5, the standard deviation is 1.393E-5 and the maximum value is 6.3E-5. But the mean of residual errors of non ground plane points is 0.00101, the standard deviation is 0.00183 and the maximum value is 0.011. The variation of residual errors of matching points is shown in (f), those less than 0.00025 are classified as ground plane points. The reprojection errors of the recovered planar homographies using our new method and several methods described by Hartley and Zisserman [2] are shown in (g), which shows us that the accuracy of the new method is very similar to the normalizing transformation method, and much better than the direct linear transform (DLT) method and the method of minimisation of symmetric transfer error. Finally, all the coplanar points and the points that lie on obstacles can be segmented using the sinusoid model. The result is shown in (b), where 'ground-plane points' and 'obstacle points' are marked as (•) and (+),respectively.

In the second experiment, the camera was moved in a forward translation mode, but in a direction not parallel to the ground. The camera axis and motion is inclined downwards towards the ground plane. In this situation, at least two matching points are needed to calculate q. We compute the FOE as $\mathbf{v} = (277.73, 302.41)$, and the vanishing line as $\mathbf{l} = (0, 0.0005, -0.13)$. Some correspondences, feature tracks, vanishing line and FOE (o) are shown in Fig. 3 (a) The segmented ground plane points (•) and obstacle points (+) are shown in Fig. 3 (b).

5.2 Experiment Using Correlations and Contour Matching

The final experiment presented in this paper, uses correlations within locally textured regions and contour matching to determine whether smooth (textureless) regions should be grouped with the ground plane. Note that an additional process is required, not described in this paper, which is our own quadtree split-merge region segmentation algorithm, which extracts homogenous regions of colour-texture. Smooth (textureless and featureless) 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 matching is easily done by 'casting' rays from the FOE recovered from *all* corner matches (note that there may be little or even no corner correspondences on the ground!).

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. The boundaries of these regions are extracted and the FOE is used to cast a ray in order to match points along corresponding boundaries. If the motion of all matched boundary points falls within the threshold of the sinusoid model in reciprocal-polar space, it is classified as belonging to the ground plane, otherwise the region is classified as an obstacle. Fig. 4 (b) 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 its boundary motion being consistent with ground plane motion in reciprocal-polar space. A second example of pixel based segmentation is shown in figures 4 (c) and (d). Note that there are some 'drop outs' in the foreground of the image, but the shape of the segmentation is excellent, to the extent that even the small black doorstop to the centre right of the original image 4 (c) has been correctly classified as an obstacle and removed in image 4 (d).

6 Conclusions

Firstly, we have presented a novel reciprocal-polar (ρ, θ) image rectification, which transforms planar image motion under pure translation into a pure shift, irrespective of the degree of perspective distortion of the planar surface. Hence correlation can be done over large translations for the class of planar homographies described as 'elations', when correlation in the original image space would fail. Secondly, we have shown that it is possible to group/segment co-planar regions by matching all co-planar pixels which have local intensity variations to the same half-wave sinusoid. This can be done without explicit computation of the homography. A nice feature of the algorithm is that *all* pixels on a plane (that have some local intensity variation) contribute to the ground plane grouping process and (if required) associated homography estimation. Thus you get a dense (i.e. pixel based) ground plane segmentation. Our results show that our algorithm performs very well to outliers and noises and the stability, accuracy and robustness performs favourably to other methods in terms of the reprojection errors of the recovered homography.

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. R Hartley and A Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge, 2001.
3. Reinhard Koch Pollefeys M and Luc Van Gool. A simple and efficient rectification method for general motion. In *ICCV*, pages 496–501, 1999.
4. Wolberg and Zokai. Robust image registration using log-polar transform. In *Proc IEEE Int. Conf. Image Processing*, 2000.