Available online at www.sciencedirect.com
SCIENCE DIRECT
Graphical Models 68 (2006) 282–293
Graphical Models
www.elsevier.com/locate/gmod

Visual reconstruction of ground plane obstacles in a sparse view robot environment

R. Laganière [a,*], H. Hajjdiab [a], A. Mitiche [b]

[a] *VIVA Research Laboratory, School of Information Technology and Engineering, University of Ottawa, Ottawa, Ont., Canada K1N 6N5*
[b] *Institut National de la Recherche Scientifique, INRS-EMT, Montreal, Que., Canada H5A 1K6*

Received 20 July 2004; received in revised form 17 October 2005; accepted 2 February 2006
Available online 31 March 2006
Communicated by Dimitris Metaxes

Abstract

The purpose of this study is to investigate a geometric/level set method to locate ground plane objects in a robot environment and reconstruct their structure from a collection of sparse views. In a first step, a model of the ground plane surface, on which the robot is operating, is obtained through the matching of the available views. This wide-baseline matching of the ground plane views allows also to compute camera pose information associated with each of these views. Based on the information obtained, reconstruction of the obstacles proceeds by minimizing an energy functional containing three terms: a term of shape-from-silhouettes consistency to characterize the ground plane objects structure and to account for possibly non-textured object surfaces; a term of visual information consistency to measure the conformity of the objects surface visual information to the acquired images; and finally, a term of regularization to bias the solution toward smooth object surfaces. The functional is minimized following the associated Euler–Lagrange surface evolution descent equations, implemented via level set PDEs to allow changes in topology while ensuring numerical stability. We provide examples of verification of the scheme on real data.

Keywords: Ground plane obstacle; 3D reconstruction; Widely separated view matching; Shape-from-silhouette; Photo-consistency; Level sets

1. Introduction

Building three-dimensional (3D) representations from a collection of two-dimensional (2D) images constitutes one of the most challenging task in computer vision [5,21]. The objective is to obtain a 3D interpretation of a scene that complies with the available observations while obeying to some constraints, these later representing prior knowledge about the observed world. A large variety of applications can benefit from the solving of this visual reconstruction problem, ranging from accurate object model building to the development of autonomous navigation systems. In this study, we propose a novel approach to determine the position and structure of ground plane objects in a robot

* Corresponding author. Fax: +1 613 562 5664.
E-mail addresses: laganier@site.uottawa.ca (R. Laganière), mitiche@inrs-emt.uquebec.ca (A. Mitiche).

1524-0703/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.gmod.2006.02.001

environment from a collection of sparse views of this environment.

When an autonomous robot is operating in a work area, it must have the ability to detect, locate, and identify the obstacles and other such objects for the purpose of avoidance, manipulation or recognition. Most often, robots move on relatively flat terrains. This fact affords a substantial simplification of the geometry of the problem and of its representation through projective relations. Reconstruction will proceed here by considering a collection of sparse views of the scene. These views can have been obtained, for example, by initially monitoring, along a peripheral path, the robot equipped with a camera to acquire images of the environment or by a team of robots collaborating together to obtain a representation of the environment. This is a very challenging situation since it requires the registration of images taken from very different points of view. But this approach offers the advantage of considerably limiting the amount of data to be processed and/or transmitted. Since many real-world robotic applications have to cope with limited bandwidth issues, this asset can be a key. In addition, 3D localization from widely separated views is generally more accurate and less sensitive to imperfect feature localization and matching due to pixelization and image noise. Following the proposed procedure, robot positions are determined, ground plane objects are detected, and, finally, their structure is reconstructed by an off-line process in charge of building and maintaining the 3D scene representation.

In general, solutions to ground plane object detection are based on the homographic relation induced by the observed planar surface. Using this relation, image points corresponding to the plane can be transferred from one view to another. In the case of a stereoscopic system of cameras, a perspective warping of the left image to the right image can be computed. Differentiating the warped and the original images leads to a rough obstacle/plane segmentation from which the obstacles can be localized [1,4,11,20]. The approach described in [7] detects and matches image points in a stereo head. Robot locations are then obtained through ground plane transformation in a way similar to the one presented in this paper. The computation of the residual disparity can also provide information about the amount of deviation of a point with respect to the reference plane [37]. When only one camera is used, information about camera motion and 3D structure of the imaged scene is generally obtained through the estimation of the optical flow field of the image sequence resulting from the robot motion [6,12,26].

The proposed reconstruction procedure is twofold. First a model of the ground plane surface, on which the robot is operating, is obtained through the matching of the available views. This wide-baseline matching of the ground plane views allows also to compute the robot (or camera) position associated with each of these views. View matching and robot localization are achieved here without using any special landmarks [30,32] or preestablished environmental map [17]. Second and based on the information obtained, reconstruction of the obstacles proceeds by minimizing an energy functional containing three terms: a term of shape-from-silhouettes consistency to characterize the ground plane objects structure and to account for possibly non-textured object surfaces; a term of visual information consistency to measure the conformity of the objects surface visual information to the acquired images; and, finally, a term of regularization to bias the solution toward smooth object surfaces. This functional is minimized following the associated Euler–Lagrange surface evolution descent equations, implemented via level set PDEs for numerical stability and to allow changes in the topology of the surface during evolution (thus handling the case of multiple obstacles).

The level set formalism for 3D reconstruction from 2D images has also been used in [8,9,36,27]. In [27], the study was concerned with temporal image sequences and short-term motion, rather than with wide baseline image sets as in this paper. In [36], 3D reconstruction of a single object and estimation of camera poses are sought under the following two assumptions: (a) the object surface is smooth and projects onto images as piecewise smooth irradiance segments with brightness discontinuities corresponding to occlusion boundaries and (b) the background and the object surface support two significantly distinct radiance functions. In [9], the method, based on perspective invariant intensity correlation, required: (a) prior accurate knowledge of camera positions and (b) pre-segmented images in terms of object and background. Their approach combines the correspondence problem and the reconstruction problem using a function which measures conformity of structure to visual data [10]. This ability to combine different constraints is one of the attractive aspects of the level set formalism; we exploit it in this study.

The methods in both [36,9] assume both a non-applicable model of brightness variations and a non-intervening background. These assumptions cannot be retained here because we are dealing with real-world scenes with objects that may not have sufficient brightness variations to inform on shape, and with backgrounds that cannot be abstracted out of the problem. The lack of sufficient object brightness variations is handled in our formulation by the shape-from-silhouette-consistency component in the energy functional, and this is a major difference with the functionals in [36,9]. The inclusion of this shape-from-silhouette in the functional is important because of the assumptions it relaxes and that must be relaxed with images of real scenes. Also, in our method, the background is not abstracted out from the reconstruction process, by assuming, for instance, a known uniform brightness background. Abstraction of the background avoids the problem where it is most delicate to solve: object boundaries. Another major difference resides in our use of robust visual information, namely color invariants rather than the raw image data, and of a geometry to account for the case of objects on a ground floor.

The remainder of this paper is organized as follows: Section 2 describes the basic models of shape-from-silhouettes consistency and of visual information consistency. Section 3 gives the formulation, the Euler–Lagrange equations and their level set expression. Section 4 gives the details for building an overhead view mosaic of the ground plane. Section 5 considers the recovery of the camera pose information. Section 6 gives experimental results of reconstruction on real scenes, and Section 7 contains a conclusion.

2. Basic models

We consider an environment consisting of an arbitrary and unknown number of objects standing at arbitrary position on a ground plane. The workplace of a robot is such an environment, objects being obstacles the robot must identify in its navigation. Our goal is to write a variational formulation to recover the position and structure of these ground plane obstacles. To achieve this goal, we have at our disposal k distinct color images of the environment, $I_j, j = 1, \ldots, k$ (as in Fig. 1). However,

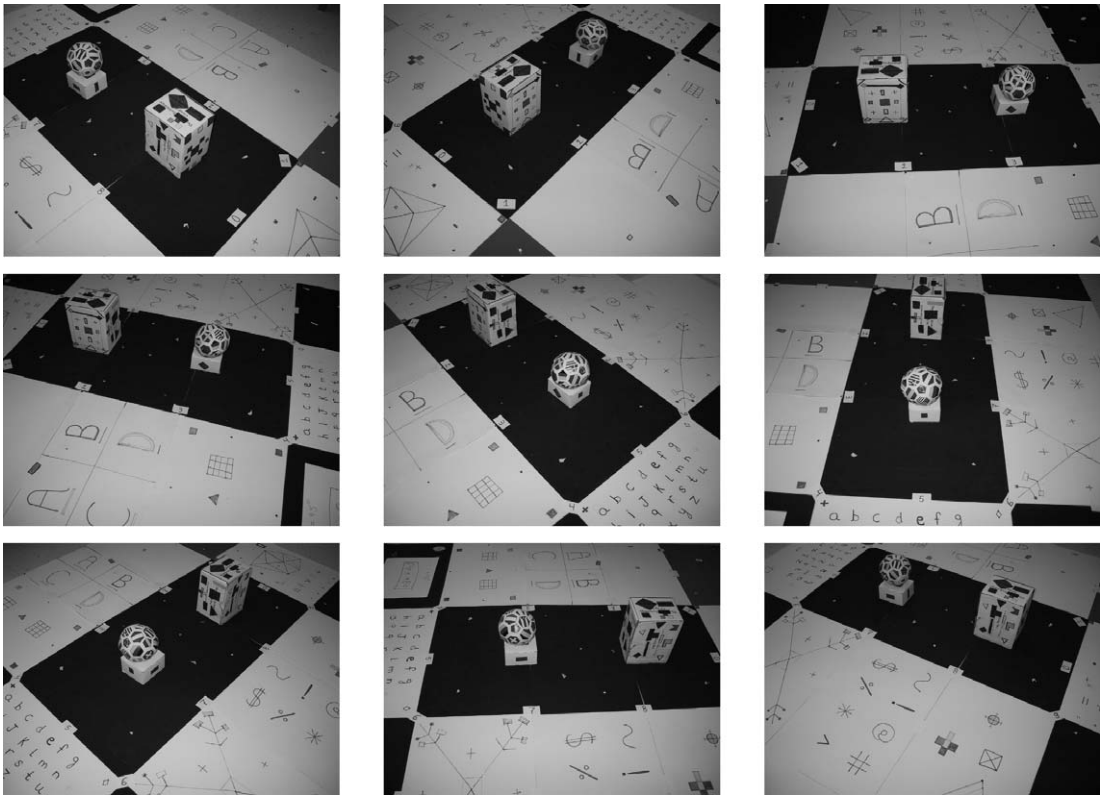

Fig. 1. Different views of a ground plane with obstacles.

because we are dealing with views taken from widely separated points of view, rather than the raw values of the acquired images I_j, $j=1,\ldots,k$, we will use normalized, Gaussian-filtered color invariants as visual information. We selected the Hilbert's color invariants, these having the advantage of requiring only first order derivatives. These derivatives are computed using Gaussian filters. Montesinos et al. [23] propose to use an invariant vector of eight such components. In our case, we just use the Gaussian-filtered color intensities, R_σ, G_σ, B_σ, and their corresponding gradient magnitudes, $|\nabla_\sigma R|$, $|\nabla_\sigma G|$, $|\nabla_\sigma B|$, with σ^2 being the variance of the Gaussian filter.

To compute the similarity between two such vectors, the components must be resized since color intensities and gradient magnitudes have different ranges of values. Moreover, since the stability of the gradient magnitude over viewpoint variation is not very good, we normalize these values using a sigmoidal function of the form:

$$s(t)=\frac{1}{1+\mathrm{e}^{-\mu(t-t_\mathrm{o})}}. \tag{1}$$

This non-linear normalization leads to a more qualitative characterization of the gradient where pixels are classified (in a fuzzy way) as points with low or high color gradient magnitude. The value t_o is the threshold that defines the point of transition between low and high magnitudes and μ controls the fuzzyness of the classification. The gradient magnitude thus transformed is similar to the measure of *edgeness* as defined in [35]. This leads us to the following normalized invariant vector:

$$I^v(X)=\left[\frac{R_\sigma}{255},\frac{G_\sigma}{255},\frac{B_\sigma}{255},s(|\nabla_\sigma R|),s(|\nabla_\sigma G|),s(|\nabla_\sigma B|)\right]^\mathrm{T}. \tag{2}$$

Euclidean distance can then be used to measure the similarity between two color image points. The use of these invariants is particularly useful at the matching step as explained in Section 4.1.

Let I^v_j, $j=1,\ldots,k$ be the k (vectorial) images of visual information. Based on projective geometry relations, we can combine these images to construct a mosaic image I_m of the ground plane (the procedure is explained in Section 4.2; an example of such mosaic is shown in Fig. 4). We can also compute the projective transformations P_j, $j=1,\ldots,k$ of the k original views. Finally, we can compute the homographic transformations $H_{\mathrm{m}j}$, $j=1,\ldots,k$ between the mosaic view and each of the original views. The details of how the mosaic view and the transformations are obtained are given in Sections 4 and 5. But assuming that all this information has been obtained, we first show here how obstacle reconstruction can proceed.

Now, given I^v_j, $j=1,\ldots,k$, I_m, P_j, $j=1,\ldots,k$, and $H_{\mathrm{m}j}$, $j=1,\ldots,k$, we can formulate the problem. The formulation incorporates two basic models: a *silhouette-consistency* model, and a *photo-consistency* model.

2.1. Silhouette-consistency model

According to the shape-from-silhouette strategy [3,38], object's visual cones from several views are intersected to obtain a visual hull within which the 3D object must lie. Each of these cones is the generalized cone defined by the union of the visual rays emanating from the view and through the object silhouette. With a large number of informative views, the object surface structure can be well approximated by the visual hull [18]. For the problem we are addressing, we cannot assume the availability of a large number of views and the visual hull one would obtain by direct application of the shape-from-silhouette strategy would be a gross estimate of the object surface structure. However, even with a small number of views, visual cones contain valuable information about object surface structure. We integrate this information in the formulation via an estimate of the probability that a given 3D point belongs to an object visual hull. The estimation of this probability is done as follows.

Let X be a 3D point. If, when viewed from camera j, the ground plane is not occluded along the visual ray through X, this visual ray does not intersect an object surface. Therefore, X is necessarily outside the visual cone of the view (when the scene contains several objects, the union of the visual cones of a view is simply called the visual cone of the view). This can be verified by comparing the visual information at the pixel of view j onto which X projects (given by $x=P_jX$) to the visual information at the pixel of the overhead mosaic which is the projection of the intersection between the visual ray and the ground plane (given by $x_\mathrm{m}=H^{-1}_{\mathrm{m}j}P_jX$), that is, by evaluating the residual:

$$r_j(X)=\|I^v_i(P_jX)-I^v_\mathrm{m}(H^{-1}_{\mathrm{m}j}P_jX)\|. \tag{3}$$

Image I^v_m is the composed mosaic image of the ground plane that is obtained from all available views. Indeed when a plane is observed from a known viewpoint, the corresponding image can be

projectively transformed into an overhead view representation; the procedure is explained in Section 4.2. The residual r_j is expected to be low for points outside the visual cone and high otherwise. Therefore, under the assumption that residuals are bounded, we express the probability that a point X is within the visual cone of view j as an increasing function of the residual, for instance:

$$P(X \in R_{\mathrm{VC}_j} | r_j(X)) \propto 1 - \mathrm{e}^{-\frac{r_j^2(X)}{\beta^2}}, \tag{4}$$

where R_{VC_j} is the region inside visual cone of view j, and $\propto$ is the proportional-to symbol. Assuming that events $X \in R_{\mathrm{VC}_i}$ and $X \in R_{\mathrm{VC}_j}$ are independent for $i \neq j$ for all X, and the visual hull being the intersection of the visual cones, the probability that a point X is within the visual hull can be expressed as:

$$P(X \in R_{\mathrm{VH}} | \{r_j\}_{j=1}^k) \propto \prod_i \left(1 - \mathrm{e}^{-\frac{r_j^2(X)}{\beta^2}}\right), \tag{5}$$

where R_{VH} is the region inside the objects visual hull. This later equation is the basis of a shape-from-silhouette approach that does not require explicit silhouette extraction; rather, it constitutes a fuzzy representation of the silhouette set to be integrated to the reconstruction process. This is in contrast with the usual schemes that rely on the availability of accurate silhouette images (often obtained by placing the object in front of a black curtain). By not explicitly imposing silhouette segmentation, introduction of errors that cannot be recovered at later stages is avoided.

2.2. Photo-consistency model

To further improve the 3D reconstruction of the observed objects, the photometric information contained in each view can be used. This can be done using a *voxel-coloring* strategy that consists in dividing the scene space into small volumetric elements (the voxels) [28]. The scene is then visited and each voxel is projected onto the input images and examined to determine if it belongs to an object surface. In summary, while silhouette-consistency refers to the region within the objects visual hull, photo-consistency concerns the objects surface.

The photo-consistency of a given 3D point can be measured by computing the standard deviation of the pixel colors to which that 3D point projects [25]. Let X be a point, and $\overline{I^v}(X)$ the visual information at X averaged over the k views then $d_j(X)$ represents the deviation from this average of the visual information at the projection of X on view j, i.e.:

$$d_j(X) = \|I_j^v(P_jX) - \overline{I^v}(X)\|. \tag{6}$$

This deviation will be low for physical points located on Lambertian surfaces because they project the same color information on all cameras. On the other hand, cameras looking at a point not located on an object surface will see different elements of the scene, thus resulting in a large visual deviation. In estimating this deviation, a more robust approach would consist in considering instead the median deviation from the pixel average color value; this would avoid having a very dissimilar pixel to excessively contribute to the consistency measure. Instead, we use a continuously derivable function, based on the idea of Geman and Reynolds [13] to use concave functions to implicitly address the problem of outliers. This is important in our application since, because of occlusion, it is not expected to have all cameras seeing all obstacle points. Our goal is simply to maximize the consensus concerning the color of an obstacle point as seen by the different views. The conformity of a given point X to the observation is then measured by:

$$h(X) = \sum_{j=1}^{k} -\mathrm{e}^{-\gamma d_j(X)}. \tag{7}$$

Such photo-consistency term will be evaluated for all voxels above the ground plane. Because consistency in photometry is observed only for those voxels lying on an object surface, this function is expected to be low for points located on an obstacle. A probabilistic formulation of photo-consistency is also proposed in [14] to iteratively identify the voxels which lie on objects' surface. The probability that a voxel exists (i.e., lie on a surface) is also estimated by Broadhurst et al. [15] in their space carving approach. The two Eqs. (5) and (7) can now be combined to obtain an energy functional representing our characterization of the ground plane objects reconstruction problem.

3. Formulation

An energy functional to be minimized can now be written. This functional will contain terms which follow from our basic models of shape-from-silhouette-consistency and visual information consistency (Eqs. (5) and (7)) and a regularization term to obtain smooth object surfaces.

3.1. Functional and Euler–Lagrange equation

Let S be a closed surface and R_S the region enclosed by S. Since the functional to be defined will be minimized, let us simply negate Eq. (5), i.e., let:

$$f(X) = \prod_i \left(e^{-\frac{r_i^2(X)}{\beta^2}} - 1 \right). \quad (8)$$

The energy functional to minimize over all closed surfaces S is:

$$E(S) = \int_{R_S} f(X)\, dV + a \int_S h(X)\, dS + b \int_S dS, \quad (9)$$

where a and b are positive constants to weigh the relative contribution of the terms in the functional. Generic derivations of the Euler–Lagrange equations corresponding to each integral in (9) (a volume integral of a scalar function and surface integrals of scalar functions) can be found in [22]. Let

$$g(X) = ah(X) + b, \quad (10)$$

Then, the Euler–Lagrange descent equation to minimize (9) is:

$$\frac{d\phi}{dt} = -(f + \nabla g \cdot \mathbf{n} - 2g\kappa)\mathbf{n}, \quad (11)$$

where ϕ is a parametrization of S, $\mathbf{n}$ is the outward unit normal on S, and κ is the mean curvature function. The right-hand side of (11) is independent of surface parametrization, as it should be, because we are interested in the image S of ϕ and not in ϕ itself.

3.2. Level set evolution equation

Execution of the descent equation by explicit representation of S as a set of points does not accommodate changes in the topology of S. An alternative execution is via level sets where S is represented implicitly as the zero level set of a one-parameter family of functions u:

$$(\forall \tau) \quad u \circ \phi(\tau)(x(\tau), y(\tau), z(\tau), \tau) = 0, \quad (12)$$

where x, y, and z are the spatial variables. Differentiation of (12) with respect to τ yields the level set equation that drives the evolution of u:

$$\nabla u \cdot \frac{d\phi}{d\tau} + \frac{\partial u}{\partial \tau} = 0. \quad (13)$$

Referring to (11) and taking u to be positive inside S_R and negative outside ($u = 0$ on S) so that $\mathbf{n} - \nabla u/\|\nabla u\|$, the evolution equation of u is, in our case:

$$\frac{\partial u}{\partial \tau} = \left(\nabla g \cdot \frac{\nabla u}{\|\nabla u\|} + 2g\kappa - f \right) \|\nabla u\|. \quad (14)$$

Because it is a gradient descent, the algorithm always converge to a local minimum. Mean curvature is expressed in terms of u as in [29]:

$$\kappa = \mathrm{div}\left(\frac{\nabla u}{\|\nabla u\|} \right). \quad (15)$$

By construction, S can be recovered at any instant as the 0-level surface of function u regardless of the changes in topology during its evolution. The basic models and the problem formulation of Sections 2 and 3 referred to a mosaic view of the ground plane and to camera pose, in particular to the projection matrices of the different views of the environment. The details of how these were obtained were not necessary to formulate the problem. We now turn our attention to these details, building a mosaic overhead view of the ground plane (Section 4) and recovery of camera pose (Section 5).

4. Building a mosaic overhead view of the ground plane

The set of images collected by the moving robot must be assembled together in order to produce a model of the ground plane over which the robot is operating. These images show the ground plane and the obstacles to be reconstructed, taken from different arbitrary and unknown viewpoints (see, for example, the images of Fig. 1). From these ones, the objective is first to build an overhead view mosaic of the ground plane. Since the images have been captured by a tilted camera of known height and tilt angle, a rough estimation of the homographic transformation between world plane (i.e., the ground plane) and each image plane can be obtained.

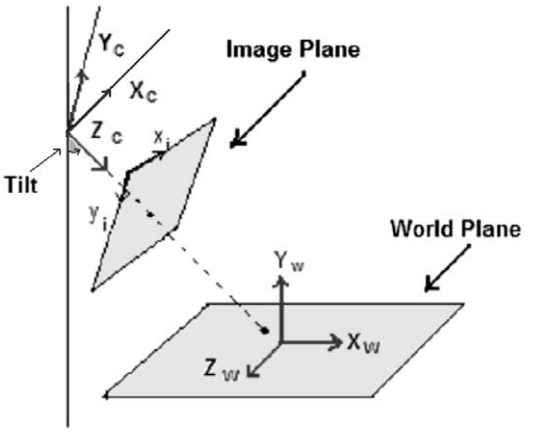

Fig. 2. The geometry of the camera system.

Fig. 2 shows the geometry of the camera system. The ground plane lies on the XZ plane and the optical axis of the camera is aligned with the Z axis. The camera, at a height h, is rotated around the X axis by an angle ϕ. When the geometry of the system is as shown, the 3×3 homography matrix H_B of the projective relation, $[x,y,1]^{\mathrm{T}} = H_B[X_{\mathrm{W}}, Z_{\mathrm{W}}, 1]^{\mathrm{T}}$, between the world plane and the corresponding image point can be described as follows:

$$H_B = \begin{bmatrix} f & u_0\cos(\phi) & u_0 h\cos(\phi) \\ 0 & f\sin(\phi)+v_0\cos(\phi) & v_0 h\cos(\phi)-fh\sin(\phi) \\ 0 & \cos(\phi) & h\cos(\phi) \end{bmatrix}, \tag{16}$$

where f represents the focal length of the camera, u_0 and v_0 are the principal point coordinates. This transformation is invertible such that the overhead view can be generated from a perspective or vice versa. Fig. 3 shows an image on which this kind of transformation has been applied.

4.1. Sparse view matching

To match two views of a scene, feature points must be detected. To this end, we used the Harris corner detector [16]. The detected feature points are then mapped on the overhead views on which matching will be performed. Indeed, working in the overhead view space has the virtue of eliminating the perspective distortion that deforms the visual patterns in each view. Consequently, to match points in these overhead views, only a rotationally invariant measure is required. Assuming that the intensity variation of images due to the changes in viewpoints is not significant, the invariant vector of Eq. (2) can give a robust characterization of the points of interest. However, because of their limited discrimination power, matching with invariants leads to several false matches. Therefore, we need to introduce an additional matching measure. The choice we made is based on the observation that while the transformation between two images of a plane is a general homography, the transformation between the two generated overhead views is an isometric transformation, i.e., it is composed of a rotation and a translation. The transformation has three degrees of freedom; two for translation and one for rotation, it can therefore be computed from two point correspondences. This isometric transformation H_S can be described as follows:

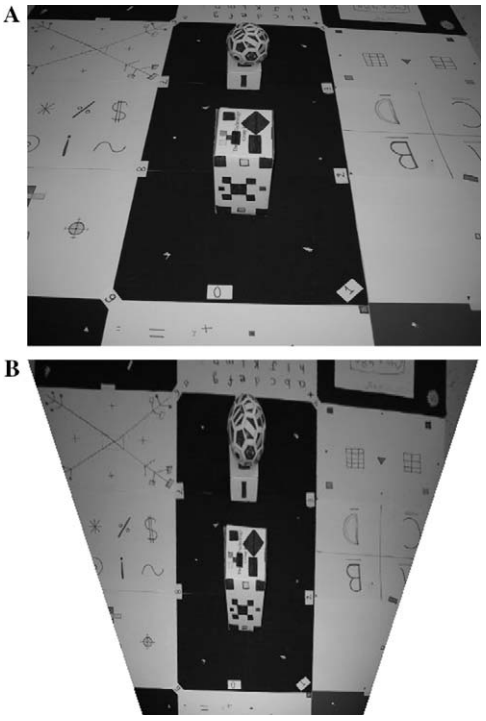

Fig. 3. (A) One additional view of the scene. (B) The generated overhead view.

$$H_S = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & T_x \\ \sin(\theta) & \cos(\theta) & T_y \\ 0 & 0 & 1 \end{bmatrix} \tag{17}$$

with θ being the angle between two cameras. An invariant in this transformation is the Euclidean distance between two points. Indeed, the line length between the two overhead views is preserved. Following a RANSAC-like scheme, we randomly select two match pairs and the length of the line segments that join the two selected points in each image is compared. If the difference in length is sufficiently low, then the points are considered to be good candidate matches. These candidate matches are then further validated by considering the intensity profiles of the candidate segments. Tell and Carlsson [31] also used a comparison between the affinely invariant Fourier features of the intensity profiles between randomly selected pairs of image interest points. We use a similar approach; however, in our algorithm, and because of the isometry that separates the two views, each line segment is simply divided into $k+1$ points. Cross correlation between the intensity profiles of the left and right lines is then computed as follows:

$$l_c = \frac{\sum_{i=0}^{k}[(A_i - \bar{A})(B_i - \bar{B})]}{\sqrt{[\sum_{i=0}^{k}(A_i - \bar{A})^2][\sum_{i=0}^{k}(B_i - \bar{B})^2]}}, \quad (18)$$

where arrays A and B contain the pixel intensity values of the $k+1$ points of the two segments (we used $k = 8$). If the correlation coefficient l_c exceeds a given value then the two segments, and therefore their corresponding end points, are assumed to match. The transformation, H_{Sij}, between the two overhead view images can then be calculated using the resulting set of matches. A best-fit finds the best isometric transformation. Using this result, the homography between the two views can then calculated by:

$$H_{ij} = H_{Bi}^{-1} H_{Sij} H_{Bj}, \quad (19)$$

where H_{Bi} is the overhead homography of view i and H_{Sij} is the overhead isometric transformation between views i and j.

To obtain a more accurate estimate of the inter-image homographies H_{ij}, the corners detected in one image are mapped using H_{ij} to the corresponding location in the other image. The neighborhood of the transferred point is searched for a corner such that the correlation difference between the two candidate corners is below a certain threshold. This updated set of matched corners is then used to refine the homography H_{ij}.

4.2. Overhead view composition

By combining the computed in-between image homographies with the overhead transformations, it is possible to build the global overhead view mosaic of the ground plane. However, when assembling the different transformed views, only the image point showing the ground plane must be used, that is, images of the obstacles must be discarded.

To do so, we use the following procedure. For each point on the ground plane, the appropriate transformation is applied to obtain the corresponding image point in each view. The mean *RGB* value is then computed and the image point that deviates the most from this mean value is discarded. This procedure is repeated until half of the image points have been discarded. The mean *RGB* value of the remaining points is then used in the mosaic composition. In our experiments, where a sufficient number of sparsely distributed views were used, this simple algorithm was able to eliminate the images of the obstacles from the ground model. Fig. 4 shows the mosaic obtained from the images of Figs. 1 and 3. Note, however, that our goal here was not to obtain a mosaic of high visual quality but rather to build a ground plane model that will be suitable in the reconstruction of the obstacles (Section 2). This is the reason why we did not apply any image enhancement algorithm (such as pixel blending).

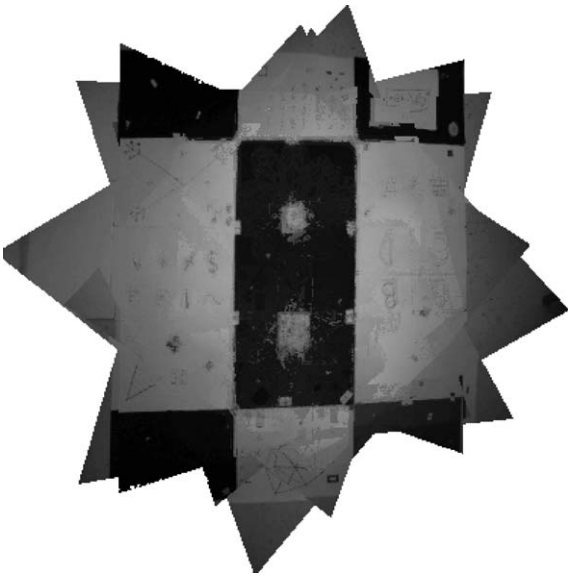

Fig. 4. The overhead view mosaic.

5. Recovery of camera pose

The computed inter-image homographies have been used to generate an overhead view mosaic representing the ground plane model. These matrices can also be used to obtain the pose of each camera. For a two-camera system, Tsai et al. [34] showed that the planar homography can be decomposed as follows:

$$H = K\left[R - \frac{Tn^{\mathrm{T}}}{d}\right]K^{-1}, \quad (20)$$

where K is the matrix containing the intrinsic parameters of the cameras, R is the rotation between the two cameras, n is the normal to the plane under consideration, and T is the translation. Finally d is the distance from the camera to the ground which can be arbitrarily set to 1. Based on this equation it is possible, as shown in [33], to extract the camera parameters through singular value decomposition of the inter-view homography. In general, the approach leads to two distinct solutions. Since in our case, the normal vector to the ground plane

is approximatively known, the correct solution can be easily identified.

6. Reconstruction experiments

Our numerical implementation is based on the level sets method proposed by Osher and Sethian in [24,29]. The first step in the implementation is to discretize the scene into 3D grid points (voxels). This is achieved by our knowledge of the ground plane equation with respect to a selected reference camera. The 3D grid points are divided into layers parallel to the ground plane. Each layer has 200×200 grid points for a total of 100 layers. The obtained 3D structures are represented in Fig. 5 where the reprojected set of voxels under a virtual light source are shown. This one has been obtained using the 10 available views of the scene (cf. Fig. 1) and the ground plane model shown in Fig. 4. The values used for the various parameters were set experimentally; however, it has been observed that a wide range of values for these parameters gave results similar to the ones shown. A total of 1500 level set iterations were required to obtain the shown result. Considering that the camera parameters (the projection matrices) have been obtained directly from the image data set (as explained in Section 5), the resulting solution succeeded well in capturing the 3D structure of the observed 'ground plane obstacle' scene.

A second result is shown in Fig. 6. It has been obtained from 12 images, some of them being shown in Fig. 7. This type of scene, i.e., non-uniform-brightness background combined with both textured and non-textured objects, is challenging to most 3D reconstruction algorithms. The evolution of the model during the level set iterative procedure is illustrated in Fig. 8 where one horizontal slice of the reconstructed model is shown at different iterations. The figure shows how, when starting

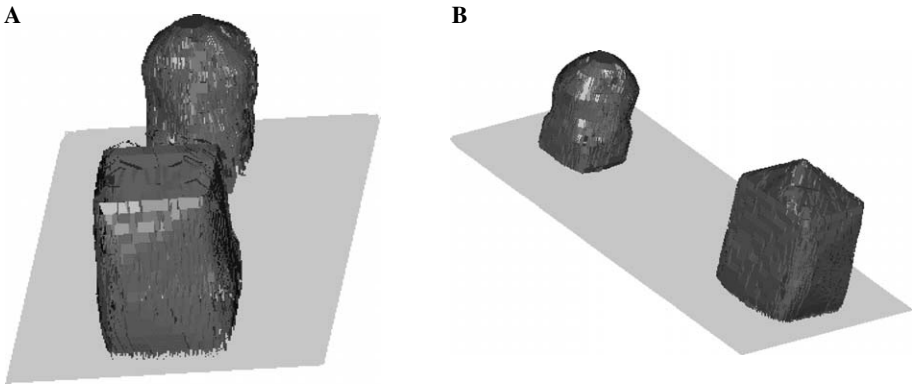

Fig. 5. The reconstructed obstacles of the scene shown in Fig. 1.

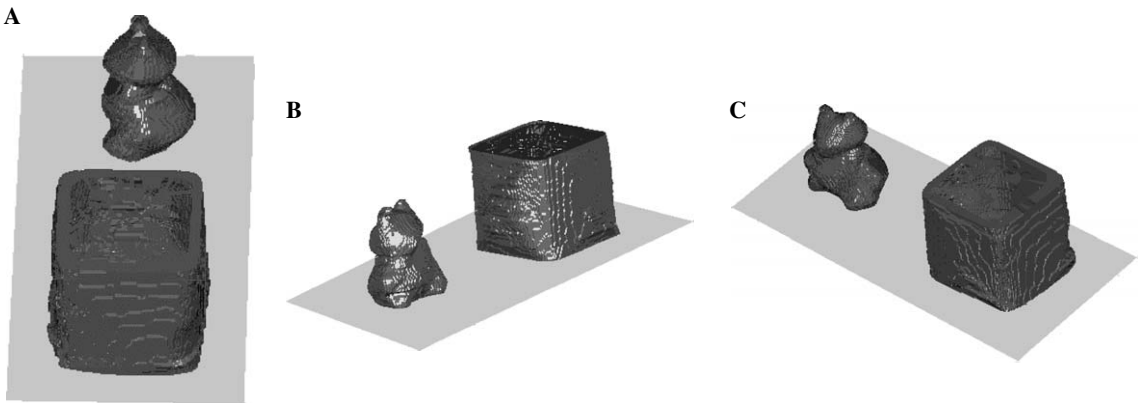

Fig. 6. The reconstructed obstacles of the scene shown in Fig. 7.

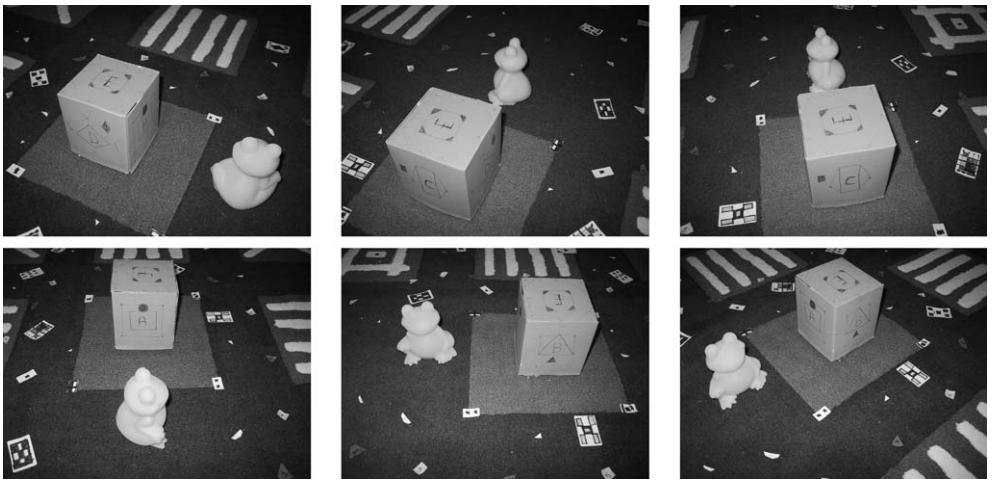

Fig. 7. A few images of a second scene.

Fig. 8. Evolution of one slice of the reconstructed scene shown in Fig. 7 after 100, 500, 900, 1100, 1200, and 1500 iterations.

with a cubic interface, the evolution can lead the level set function to break into several distinct volumes some of them eventually vanishing if not supported by a sufficient level of energy. It should also be noted that level set evolution is a time consuming process. However, this is not critical in the present application because reconstruction would be performed off-line, most probably by a base station, upon reception of the data collected by the robot.

Finally, we also test our algorithm on a natural outdoor scene. A total of 14 images were used in this case. Fig. 9(A) shows one of these images and the resulting 3D structure is represented in Fig. 9(B).

7. Conclusion

This study addressed the problem of reconstructing the structure of objects on the ground plane of a

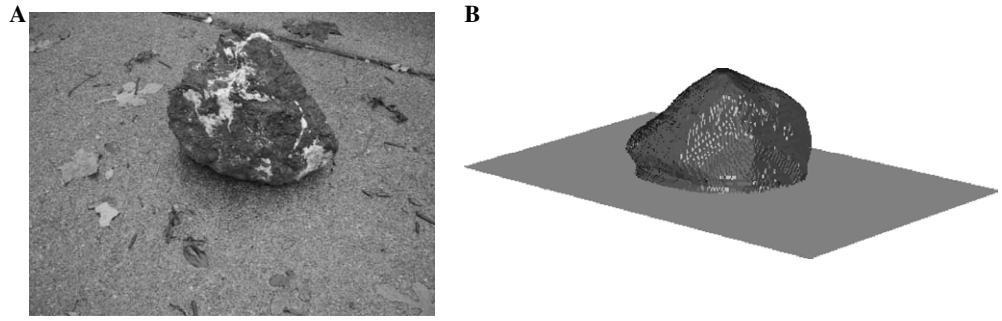

Fig. 9. One image of an outdoor scene and the reconstructed obstacle.

robot environment. The proposed solution is based on a general, variational statement of the problem, solved via level set PDEs. The objective energy functional to be minimized includes a novel term (silhouette-consistency) in addition to the terms of conformity to data (expressed, here, as a voxel coloring problem), and regularization. This extends the problem statement to (an arbitrary number of) real-world objects which may lack the brightness variations necessary to infer a shape-from-brightness process.

Different factors can affect the accuracy of the resulting 3D model. The precision of the estimated camera positions is of prime importance. This positional information is obtained here from the matching of natural features. When coupled with refining techniques, such as bundle adjustment, very accurate projection matrices can be obtained, as long as sufficient features are available in the scene. The set of available views must also ensure a complete coverage of the scene. The formulation also requires a good level of redundancy between the views; this is generally achieved by a 20–30° view separation, i.e., a loop of 10–20 views. Accuracy of the produced 3D model is ultimately determined by the resolution of the 3D grid used for level set evolution. The number of voxels in the grid has also an important impact on the required computational effort. The use of multi-resolution strategies (e.g., octree representations) to accelerate level set evolution can improve the efficiency of the algorithm. Faster implementations of level set methods are the object of active research.

Finally, the solution proposed also accounts for real-world robot environment image data. This context required the use of image invariants for sparse view matching, the estimation of camera pose from ground plane geometry, and of the construction of a ground plane mosaic model. The result is a 3D representation of the environment explored by a mobile robot, that is inferred strictly from the visual data and that has an accuracy largely sufficient for the robot to perform tasks such as avoidance, localization, recognition, and manipulation.

References

[1] P. Batavia, S. Singh, Obstacle detection using color segmentation and color stereo homography, in: IEEE Conference on Robotics and Automation, Seoul Korea (2001).
[3] C. Chian, J. Aggarwal, Model reconstruction and shape recognition form occluded contours, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 372–389.
[4] Y. Chow, R. Chung, Obstacle avoidance of legged robot without 3d reconstruction of the surroundings, in: Proc. of the IEEE Conf. on Robotics and Automation, 2000, pp. 2316–2321.
[5] U.R. Dhond, J.K. Aggarwal, Structure from stereo—a review, IEEE Trans. Syst. Man Cyb. 19 (1989) 1489–1510.
[6] W. Enkelmann, Obstacle detection by evaluation of optical flow fields from image sequences, Image Mach. Comput. 9 (3) (1991) 160–168.
[7] P.A. Beardsley, P.A. Reid, A. Zisserman, D.W. Murray, Active visual navigation using non-metric structure, Int. Conf. Comput. Vis. (1995) 58–64.
[8] V. Caselles, R. Kimmel, G. Sapiro, C. Sbert, 3d active contours, Int. Conf. Anal. Opt. Syst. (1996) 43–49.
[9] O. Faugeras, R. Keriven, Variational principles, surface evolution, PDE's, level set methods and the stereo problem, IEEE T. Image Process. 7 (3) (1998) 336–344.
[10] O. Faugeras, J. Gomes, R. Keriven, Computational stereo: a variational method, in: S. Osher, N. Paragios (Eds.), Geometric Level Set methods in Imaging Vision and Graphics, Springer Verlag, Berlin, 2003, p. 532.
[11] F. Ferrari, E. Grosso, G. Sandini, M. Magassi, A stereo vision system for real time obstacle avoidance in unknown environment, IEEE Int. Workshop on Intelligent Robots and Systems, 1990, pp. 703–708.
[12] P. Fornland, Direct obstacle detection and motion from spatio-temporal derivatives, in: CAIP, Prague, Czech Republic, Sept.1995.
[13] D. Geman, G. Reynolds, Constrained restoration and the recovery of discontinuities, IEEE T. Pattern Anal. Mach. Intell. 14 (3) (1992) 367–383.

[14] M. Agrawal, L.S. Davis, A probabilistic framework for surface reconstruction from multiple images, IEEE Conf. Comput. Vision Pattern Recogn. 2 (2001) 470–476.
[15] A. Broadhurst, T.W. Drummund, R. Cipolla, A probabilistic framework for space carving, Int. Conf. Comput. Vision 1 (2001) 7–14.
[16] C. Harris, M. Stephens, A combined corner and edge detector, Alvey Vision Conf. (1988) 147–151.
[17] E. Krotkov, Mobile robot localization using single image, in: Proc. IEEE Int. Conf. on Robotics and Automation, 1989, pp. 978–983.
[18] A. Laurentini, The visual hull concept for silhouette-based image understanding, IEEE Trans. Pattern Anal. Mach. Intell. 17 (2) (1997) 150–162.
[20] M. Lourakis, S. Orphanoudakis, Visual detection of obstacles assuming a locally planar ground, Proc. 3rd Asian Conf. on Computer Vision 2 (1998) 527–534.
[21] A. Mitiche, Computational Analysis of Visual Motion, Plenum Press, 1994.
[22] A. Mitiche, R. Feghali, A. Mansouri, Motion tracking as spatio-temporal motion boundary detection, Robot. Auton. Syst. 43 (2003) 39–50.
[23] P. Montesinos, V. Gouet, R. Deriche, D. Pel, Matching color uncalibrated images using differential invariants, Image Vision Comput. 18 (9) (2000) 659–672.
[24] S. Osher, J. Sethian, Fronts propagation with curvature-dependant speed: algorithms based on hamilton–jacobi equations, J. Comp. Phys. (79) (1988) 12–49.
[25] A.C. Prock, C.R. Dyer, Towards real-time voxel coloring, Proc. Image Understanding Workshop, 1998), pp. 315–321.
[26] J. Santos-Victor, G. Sandini, Uncalibrated obstacle detection using normal flow, Mach. Vision Appl. (1996) 130–137.
[27] H. Sekatti, A. Mitiche, Dense 3D interpretation of image sequences: a variational approach using anisotropic diffusion, in: IAPR International Conference on Image Analysis and Processing, Montova, Italy, 2003.
[28] S. Seitz, C. Dyer, Photorealistic scene reconstruction by voxel coloring, in: IEEE Conf. on Computer Vision and Pattern Recognition, 1997, pp. 1067–1073.
[29] J. Sethian, Level Set Methods and Fast Marching Methods, Cambridge University Press, Cambridge, 1996.
[30] R. Sim, G. Dudek, Mobile robot localization from learned landmarks, in: Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems (IROS), 1998.
[31] D. Tell, S. Carlsson, Wide baseline point matching using affine invariants computed from intensity profiles, European Conf. on Computer Vision, 2000, pp. 814–828.
[32] S. Thrun, Y. Liu, Multi-robot SLAM with sparse extended information filers, in: Proceedings of the 11th International Symposium of Robotics Research (ISRR'03), 2003.
[33] B. Triggs, Auto-calibration from planar scenes, European Conf. on Computer Vision, 1998, pp. 89–105.
[34] R.Y. Tsai, T. Huang, W. Zhu, Estimating three-dimensional motion parameters of a rigid planar patch, IEEE Acoustic Speech Signal Process. 30 (4) (1982) 525–534.
[35] J. Weng, N. Ahuja, T. Huang, Two-view matching, Int. Conf. on Computer Vision, 1988, pp. 65–73.
[36] A. Yezzi, S. Soatto, Structure from motion for scenes without features, in: Proceeding of the IEEE Conf. on Comp. Vis. and Patt. Recog, 2003.
[37] Z. Zhang, R. Weiss, A. Hanson, Qualitative obstacle detection, in: IEEE Conf. on Computer Vision and Pattern Recognition, 1994, pp. 554–559.
[38] J. Zheng, Acquiring 3d models from sequences of contours, IEEE Trans. Pattern Anal. Mach. Intell. 16 (2) (1994) 163–178.