Registration and interactive planar segmentation for stereo images of polyhedral scenes

Javier Flavio Vigueras\(^1\) and Mariano Rivera\(^1\)
\(^1\)Center for Mathematical Research
Centro de Investigacion en Matematicas (CIMAT)
Jalisco S/N, Colonia Valenciana
36240, Guanajuato, Gto, Mexico

Abstract

We introduce a two-step iterative segmentation and registration method to find coplanar surfaces among stereo images of a polyhedral environment. The novelties of this paper are: (i) to propose a user-defined initialization easing the image matching and segmentation, (ii) to incorporate color appearance and planar projection information into a Bayesian segmentation scheme, and (iii) to add consistency to the projective transformations related to the polyhedral structure of the scenes. The method utilizes an assisted Bayesian color segmentation scheme. The initial user-assisted segmentation is used to define search regions for planar homography image registering. The two reliable methods cooperate together to obtain probabilities for coplanar regions with similar color information that are used to get a new segmentation by means of Quadratic Markov Measure Fields (QMMF). We search for the best regions by iterating both steps: registration and segmentation.

Key words: interactive computer vision, stereo, registering, segmentation, coplanarity, color

1 Introduction

Planar surfaces are often found in nature as well as in artificial manmade environments: outdoor scenes are commonly formed by polyhedral buildings; indoor scenes contain floors, walls, desks, etc. Planes have a constrained representation and ease various computer vision tasks such as camera calibration [20, 5], camera localization [7, 17], robot navigation [13], and 3D reconstruction [4, 22]. Plane-based algorithms are commonly stable but they may become ill-conditioned when they are applied to wrong coplanar features, and therefore
it is very important to know which image regions correspond to individual planes. By these reason, several works have been conducted on plane detection and segmentation.

1.1 Previous works

Unsupervised plane detection and segmentation are commonly solved using sparse image key-points (using structure from motion techniques [18,17,2]), disparity maps [8,21], optical flow approaches [25], triangular surfaces [11] or range images [23]. Nevertheless, many of these approaches require to perform a 3D reconstruction; other methods assume that the plane is mostly textured [2], that a single plane is dominant in the image, or the camera require a rough calibration [1,21], constraining the range of application of those methods.

Image acquisition from a moving camera over a polyhedral scene imposes known constraints on matching information for every couple of views and it is possible to extract planar segmentation from them without explicitly performing a 3D reconstruction [24]. Few approaches have tried to conduct segmentation on disparity models [19] but they often fail because several ambiguities arise on considering disparity information alone. In Figure 1 images of outdoor examples are shown. Matching sparse features often fails in cases like the tower (first column) and the roof (second column) because both planes have mostly the same texture. On the other hand, untextured surfaces (like the three walls in the third column figures) are also difficult to distinguish as different planes for optical flow techniques, disparity maps and feature-based methods, because they are usually mistaken or blended.

Several automatic solutions (cited in [16]) have been developed for the planar segmentation problem from stereo views. Nevertheless, correct results are not reachable under some circumstances (described below in this paper), and only few automatic approaches may attain satisfactory results but in a unmanageable amount of time for some applications [16,9]. On the other hand, humans
are capable to correctly distinguish distinct planar surfaces from a single image in a very short time, but there are two main disadvantages on doing this task as a fully manual procedure: (i) a completely user-assisted segmentation may be tedious when the borders of planar surfaces are not clearly distinguished into the image, when these borders have complex shapes, or when the interfaces are not friendly enough; (ii) for some tasks, it is necessary to obtain metric information from the images corresponding to the featured planes and, even if the user may satisfactorily segment the planar surfaces, it is required a second phase anyway in order to get parametric data from user’s defined regions.

1.2 Proposed approach

Several computer vision methods seem to be very sensitive to the initial estimations required as input and, for many applications, choosing the adequate initial values may become an important or tedious problem itself. Interactive approaches have allowed to separate clearly the initialization stage from the automatic image processing, and improve the development of efficient computer vision tools due to the inclusion of knowledge given by human experts.

In order to cope with the problems stated in 1.1, we propose a user-assisted segmentation method combining color information and motion matchings of observed coplanar features in a two-image set. The novelty of this paper is to directly compute both homography parameters and dense region segmentation by means of a brief interaction with the user, instead of unsupervised state-of-the-art’ methods that generally consider either case: (i) finding sparse coplanar points and then fitting a planar surface (implying occlusion and convexity problems), or (ii) computing general optical flow or disparity maps and later trying to fit a planar surface.

Our approach is based on a rough solution given into an initial stage by the user and an iterative two-step algorithm: (1) registering two views of the scene in order to find the corresponding planar homographies for every plane and (2) segmenting the image using a Bayesian segmentation approach. We refine the homography models using the new marginal a posteriori probabilities obtained from segmentation as a registering mask, and repeat the two-step procedure until convergence. As input for the planar segmentation step, model likelihoods are estimated by combining planar homographies fitting between the two views and their corresponding color information.

The paper is organized as follows: our scheme for simultaneous segmentation and image registration for multi-planar environments is given in section 2. Next sections present the theoretical framework for parametric image reg-
istration (section 3), likelihood computation (section 4) and the Quadratic Markov Measure Fields (QMMF) approach for Bayesian segmentation (section 5). Section 6 presents a framework for computing homographies with polyhedral consistency. Section 7 exhibit results for some examples. Finally, a discussion on the method is summarized in section 8.

2 Overview

The problem issued in this paper may be formulated as follows: given some user-defined samples of coplanar regions taken from uncalibrated stereo views of a scene, we aim to estimate the corresponding planar projective transformations (homographies) and to extend the coplanar regions by segmenting the 2D stereo images. The vision system is modeled using the classical pin-hole camera model, which intrinsic parameters are supposed unknown and not necessarily constant. Before describing the system components in detail, we provide an overview of our method:

In our approach, the only required interaction is the user marking some samples of coplanar regions over one of the images at the procedure’s beginning. Such a simple interaction provides very important information: the numbers of planes in the scene and small regions that undoubtedly belongs to each plane. Then, a Bayesian approach based on QMMF (see section 5) is applied in order to estimate the marginal probabilities based on color information around the user-defined sample regions. Once the coplanar regions were defined and the color probability fields were obtained, fully automated image registration and segmentation are iteratively done.

The core of our algorithm is an energy function relating the model regions and the respective homographies, an iterated minimization allows us:

(1) to compute the homography coefficients corresponding to each identified region, and thus matching region intensities between both images;
(2) to compute new marginal probabilities that take into account spatial coherence. At the end of this stage, new a posteriori probabilities involving color and planar information are used to redefine registration masks for step 1.

The main steps of this algorithm are described in table 1.
(1) Acquire two images of a planar piecewise environment.

(2) At one image, the user selects samples of the coplanar regions (see Figures 3-a, 8-a and 8-c) to be used in the registering process.

(3) Estimate the marginal probabilities (by QMMF, section 5) corresponding to the color information of the user-defined regions (section 4.1).

(4) Step loop computation of the registration parameters (homography) and the segmented regions:
   (a) Registration (sections 3) from coarse scale to finer scales:
      • Establish the energy function for every defined region (section 3 for the original function and section 6 for the consistent approach).
      • Compute the coefficients of each planar homography by minimizing the corresponding energy function.
      • Update parameters for the multi-scale scheme.
   (b) Segmentation:
      • Compute the move matching likelihoods for planar homographies (section 4.2).
      • Compute the composed likelihoods by the product of move matching likelihoods and color probabilities (section 4.3).
      • Estimate the composed marginal probabilities by QMMF Bayesian segmentation (section 5).
      • For each model, select the biggest connected region with marginal probabilities greater than 0.5 (using a fill method). These regions will be used at the registration step instead of the initial color probability samples.

(5) (Optional) Refine region borders by using intersections between planes (section 6.3).

Table 1
Overview of the method.

3 Image registration

3.1 Energy function for parametric registration

Standard intensity-based approaches are founded on the luminance (chromance) constancy condition given as \( I_2(T(\mathbf{x}, \Theta)) = I_1(\mathbf{x}) + \eta(\mathbf{x}) \) where \( I_1 \) and \( I_2 \) are the gray-level (color) intensities for the first and the second views, respectively, at a given image location \( \mathbf{x} \) in the image lattice \( \mathcal{L} \), \( T \) is the parametric transformation (with parameters \( \Theta \)) describing the displacement of a 2D point in the first image to its corresponding projection at the second image, and \( \eta \) is identically distributed image noise for gray-level intensities.
Before to present the whole QMMF based energy function, we focus our attention in the image registering stage. Our registration is based on a cost function written as the sum of squared differences of the above intensities given as follows:

$$E(\Theta) = \sum_{x \in L} (I_2(T(x, \Theta)) - I_1(x))^2.$$  

(1)

Such a cost function will be useful for defining the QMMF energy.

### 3.2 Parametric registration models

**Camera model.** In this paper, we consider that the camera model is the pinhole (perspective projection) model, which associates a point $X$ in the scene to a point $x$ in the image by $\tilde{x} \sim PX$, where $P$ is a $3 \times 4$ matrix called the projection matrix, $x$ and $X$ are 2D and 3D points, respectively, expressed in homogeneous coordinates: $\tilde{x} = (x, 1)^T$ and $\tilde{X} = (X, 1)^T$. $a \sim b$ means that $a$ and $b$ are equivalent up to a scale factor, thus, there exists a scalar $\lambda$ such that $a = \lambda b$.

**Planar projective transformations (homographies).** Let us now constrain $X$ to lie on a plane $\pi$ (see Figure 2). If $x_i = (x_i, y_i, 1)$ and $x'_i = (x'_i, y'_i, 1)$ are two projections of the same 3D point $X$ on the space for the first and the second image respectively, the projections are related by $x'_i \sim Hx_i$, then, the transformation modeling the 2D movement of coplanar points under perspective projection is given by a $3 \times 3$ homography matrix $H = \{h_{ij}|i, j = 1, 2, 3\}$.

For each observed plane, a homography is implicitly related to the cameras relative position, the plane position and the cameras projective properties [6]. Nevertheless, the coefficients of $H$ may be directly estimated from image correspondences and it is not necessary to know explicitly the camera and plane parameters.
**Affine model.** Another very common transformation used for image registering is the affine model \((h_{31} = h_{32} = 0, h_{33} = 1)\), although this model is not always capable to describe the correspondences due to perspective projections. The affine model may be used when the image planes for both cameras lie on the same plane; it also represents a good approximation if both cameras have large focal lengths and the distance between the scene and the cameras is very large.

### 3.3 Retrieving the homographies

By combining the cost function given by Equation (1) and the parametric transformation models (section 3.2), we get:

\[
E(H_k) = \sum_{x \in \mathcal{L}} \left[ w_k(x) \| I_2(x') - I_1(x) \|^2 \right]
\]

(2)

where

\[
x' = z(H_k, x, y) \begin{pmatrix} h_{11}^{(k)} x + h_{12}^{(k)} y + h_{13}^{(k)} \\ h_{21}^{(k)} x + h_{22}^{(k)} y + h_{23}^{(k)} \end{pmatrix}
\]

\(N\) is the number of calibration planes, \(H_k\) is the transformation matrix for model \(k\), \(w_k(x) \in [0, 1]\) acts as a membership variable for the \(k\)-th planar region (the computation of such memberships is addressed in next section); the perspective division quotient is \(z(H_k, x, y) = (h_{31}^{(k)} x + h_{32}^{(k)} y + h_{33}^{(k)})^{-1}\) for the homography model and \(z(H_k, x, y) = 1\) for the affine model.

The homography induced by each plane is computed by minimizing each energy function \(E(H_k)\). A non-linear optimization method should be used to minimize these functions; in our implementation we opted for using a Levenberg-Marquardt technique [12].

In order to avoid converging to local minima, we propose to use a multi-scale approach on matching both images, as it has already been used in optical flow estimation: firstly, the image is sub-sampled several levels and a homography is estimated for a coarse scale; the coarse scale solution should be then used as initial parameters for solving the registration problem at a finer scale.
4 Likelihood estimation

4.1 Likelihood associated to color information

Let $I$ be an image such that $I(x) \in t$, where $t = \{t_1, t_2, \ldots, t_T\}$, are the pixel values, then the density distribution for the color classes are empirically estimated by using a histogram technique. That is, by smoothing the histograms such that $g_{ki}$ is the number of hand labeled pixels with value $t_i$ for the class $k$, then the normalized histogram is computed with

$$\hat{g}_{ki} = \frac{g_{ki}}{\sum_{j=1}^{N} g_{kj}}.$$  \hspace{1cm} (3)

The likelihood function (LF) is then computed with:

$$V_k^I = \frac{\hat{g}_{ki} + \epsilon}{\sum_{j=1}^{N}(\hat{g}_{ji} + \epsilon)} \quad \forall k$$  \hspace{1cm} (4)

where $\epsilon$ is a small constant (e.g. $1 \times 10^{-8}$). Thus, the likelihood of the pixel $x$ for a given class $k$ is computed with:

$$v_k^C(x) = V_{k,g(x)}.$$  \hspace{1cm} (5)

4.2 Likelihood associated to coplanarity

In order to compute the likelihood of a pixel image associated to a homography matching model between two images, it is necessary to establish a probabilistic measure to compare two pixels information. The likelihood of a pixel associated to a homography transformation can be estimated as the probability that the given pixel is translated, at the second image, to another pixel with similar color information.

We assume that the probability distribution for noise $\eta^{(C)}$ on every color channel $A \in \{R, G, B\}$ is Gaussian with mean zero and standard deviation $\sigma$:

$$P(\eta^{(A)}(z)) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{z^2}{2\sigma^2} \right)$$

Then, taken into account only one-to-one pixel correspondences, the likelihood is stated as:

$$v_k^P(x) = \prod_{a \in \{R,G,B\}} P \left( \eta^{(a)} \left( I_1^{(a)}(x) - I_2^{(a)} (T(x,H_k)) \right) \right)$$  \hspace{1cm} (6)
Likelihoods considering neighborhood texture information around $x$ can be computed over a window of a given size ($W_x$, e.g. $3 \times 3$ pixels) by:

$$v_k^H(x) = \prod_{y \in W_x} v_k^P(y).$$

(7)

4.3 Composed likelihood: coplanarity and color

Trying to distinguish coplanar regions only through the planarity likelihood is not always a solvable problem, because a pixel projected by a homography into the second image may fall into a region with a uniform color (e.g. the sky, untextured walls) and, in this case, the coplanarity score could be high even if the planar model is not correct. It means that the coplanarity likelihood is only informative at non-flat regions, mainly in high-gradient regions.

In order to propose a more informative criterion, we estimate a joint likelihood between planar and color information for each plane, such that the composed likelihood is computed as the product of Equation (5) and Equation (7):

$$v_k(x) = v_k^C(x)v_k^H(x)$$

(8)

5 Quadratic Markov Measure Fields (QMMF)

5.1 Computation of the Measure Probability Field

Given a set of likelihoods containing information about plane and color similarity to a given model, the goal is to find a probability field $p$ indicating which model is supported for every pixel in the image. In particular probability $p_k(x)$ for pixel $x$ will be the highest among $p = \{p_i(x) : i = 1, \ldots, N\}$ if $x \in R_k$, where $R_k$ is the region in the image that corresponds to model $k$, and $N$ is the number of models. As $p$ is a probability measure field, it has to satisfy the constraints:

$$\sum_{k=1}^{N} p_k(x) = 1 \quad \text{and} \quad p_k(x) \geq 0 \quad \forall k, x$$

(9)

According to the QMMF framework proposed in [15], the optimal estimator for the probability measure field is the minimum of the function:
\[ U(p) = \sum_{x \in L} \left\{ \sum_{k=1}^{N} \left[ -\log (v_k(x)) - \mu \right] p_k^2(x) + \lambda \sum_{y \in N_x} \|p(x) - p(y)\|^2 \right\} \quad (10) \]

subject to the constraints (9). In Equation (10), \( v(x) = (v_1(x), \ldots, v_N(x)) \) is the likelihood vector with:

\[ v_k(x) = P(I(x)|x \in R_k, \theta_k) \]

\( \theta_k \) are the parameters for model \( k \), \( \lambda \) is a non-negative parameter controlling spatial coherence between all the neighbors pixels \( (x, y) \), and the parameter \( \mu \) is added to control the entropy of the probability field.

Gauss-Seidel method may be used for the minimization of the function given by Equation (10) using the iterative rule [15]:

\[ p_k(x) = \frac{n_k(x)}{m_k(x)} + \frac{1 - \sum_{i=1}^{N} \frac{n_i(x)}{m_i(x)}}{m_k(x) \sum_{i=1}^{N} \frac{1}{m_i(x)}} \]

where:

\[ n_k(x) = \lambda \sum_{y \in N_x} p_k(y) \]

\[ m_k(x) = -\log v_k(x) - \mu + \lambda \#(N_x) \]

To satisfy the constraints (9), all negative values should be set to zero and renormalization must be done if needed.

**Relation with HMMF.** Hidden Markov Measure Fields is a Bayesian segmentation scheme strongly related to QMMF. In [15], it is shown that for vectors \( p(x) \) with low entropy, the optimal HMMF estimator is approximated by the QMMF solution.

**Binary QMMF.** The general QMMF approach should satisfy the constraints (9). Nevertheless, the sum of probabilities is one only if every pixel at the image corresponds to an observation model. In the planar case, it means that all the observed surfaces should be planar (or approximated by planes) and that there would be a plane-color model corresponding to it. In other case, some pixels may be misclassified.

In the cases where just a few number of planes were required or the image is not completely formed by plane surfaces, we use a QMMF approach for binary classification [14]. Instead of computing simultaneously the probability field for \( N \) models, for each given model, we segment the image in two regions: one belonging to this model and another one for pixels that do not correspond.
This approach will conduct $N$ binary QMMF segmentation procedures and is faster than the general QMMF method:

$$
U'(p_k) = \sum_{x \in L} \left\{ \left[ - \log (\hat{v}_k(x)) - \mu \right] p_k^2(x) \\
+ \left[ - \log (\hat{u}_k(x)) - \mu \right] (1 - p_k(x))^2 \\
+ \lambda \sum_{y \in N_x} (p_k(x) - p_k(y))^2 \right\}
$$

(11)

where the input likelihoods for each binary segmentation are computed as follows:

$$
\hat{v}_k(x) = \frac{v_k(x) + \epsilon}{\sum_{j=1}^N (v_j(x) + \epsilon)} \quad \forall k = 1, \ldots, N
$$

$$
\hat{u}_k(x) = \max \left\{ \hat{v}_i(x) | i \neq k \right\}
$$

(12)

5.2 Model parameter computation

The QMMF model allows to estimate the likelihood function (12) parameters [15]. To make explicit the parameter dependency of the QMMF energy functional (10), the first term is expressed as:

$$
\sum_{x \in L} \sum_{k=1}^N \left\{ p_k^2(x) \left[ \sum_{y \in W_x} \| I_2(T(y, \Theta)) - I_1(y) \|^2 - \log \left( v_k^C(x) \right) \right] \right\}
$$

by neglecting independent terms in $\Theta$:

$$
U(\Theta) = \sum_{x \in L} \sum_{k=1}^N \left\{ w_k(x) \| I_2(T(x, \Theta)) - I_1(x) \|^2 \right\}
$$

where $w_k(x) = \sum_{y \in W_x} p_k^2(y)$. This energy (13) is similar to $E(H_k)$ in (2).

6 Our approach of projective representation for polyhedral structures

6.1 Dependency between homographies

In section 3, we have proposed a registration method based on the minimization of a function expressed in terms of the coefficients of each homography transformation $H_k$. Nevertheless, this basic method does not take into account
the fact that the set of homographies associated to different planes observed in both views are not independent, and the results may be not consistent with the epipolar geometry of the camera array, as most of the multiplanar techniques found in the literature \[1,2,9,13,19,21,24,25\].

Actually, each planar homography is associated to the cameras intrinsic parameters, relative motion between both views, and plane equations, as follows \[6\]:

\[
H_k \sim K_2 \left( R - tv_k^T \right) K_1^{-1}
\]

where \(K_1\) and \(K_2\) are the intrinsic parameters matrix for the first and the second cameras, respectively; \(R\) and \(t\) are the relative rotation and the relative translation between both views, resp.; and \(v_k\) is a vector associated to the \(k\)-th plane, such that the plane equation is \(v_k \cdot x + 1 = 0\). The motion and intrinsic parameters are common for all the possible planar homographies and the only different terms between these homographies are related with vectors \(v_k\).

Although there exist techniques to recover the whole set of parameters (intrinsic, motion and plane equations), these methods require a large number of either planes or views of the scene in order to obtain stable results. But auto-calibration and structure from motion are not necessary to establish a coherence for the homography set. In Ref. \[3\], the authors proposed a framework relating two planar homographies associated to a given epipolar geometry. By using a similar approach, but avoiding the fundamental matrix computation and the stability problems associated to its estimation, we define a set of matrices \(M_{ij}\) such that \(M_{ij} = H_i^{-1}H_j\) for every couple of observed planes. Thus,

\[
M_{ij} \sim K_1 \left( R - tv_i^T \right)^{-1} \left( R - tv_j^T \right) K_1^{-1}.
\]

Using the Sherman-Morrison formula \[6\], we obtain:

\[
M_{ij} \sim K_1 \left( R^{-1} + \frac{R^{-1}tv_i^TR^{-1}}{I - v_i^TR^{-1}t} \right) \left( R - tv_j^T \right) K_1^{-1}
\sim K_1 \left( I + R^{-1}t (v_i - v_j)^T \right) K_1^{-1}
\sim I + \left[ K_1R^{-1}t \right] \left[ K_1^{-T}(v_i - v_j) \right]^T
\]

It means that \(M_{ij} = I + es_{ij}^T\), where \(s_{ij} = K_1^{-T}(v_i - v_j)\) and \(e = K_1R^{-1}t\) is the right epipole of the given two-view geometry. Hence, every homography may be written by:

\[
H_j \sim H_iM_{ij} = H_i \left( I + es_{ij}^T \right)
\]

Consequently, we may express every planar homography in terms of a reference homography \(H_{ref}\), the epipole position \(e\) and a 3D vector \(s_{ij}\) associated to the
position between planes projection $i$ and $j$. The reference homography can be any plane homography and the consistent homography estimation transforms every independent function optimization (Equation 2) into the minimization of a unique function:

$$E(H_{ref}, e, s_{ref,1}, \cdots, s_{ref,N}) = \sum_{k=1}^{N} \sum_{x \in L} \left[ w_k(x) \| I_2(x_k') - I_1(x) \|^2 \right]$$

(14)

where $x_k' \sim H_{ref} \left( I + e s_{ref,k}' \right) x = H_{ref} \left( x + (s_{ref,k} \cdot e) \right)$. The minimization step can be conducted by means of the Levenberg-Marquardt method that seems to be very stable for these functions.

Our method has two main advantages respect to the approach presented in Ref. [3]: (i) it is not necessary to explicitly estimate the epipolar geometry (fundamental matrix) for the camera array, which may be very unstable for some scene structures or camera movements, and (ii) the parametrization does not depend on the infinite or infinite nature of the epipoles, that actually changes the mathematical form of the homography dependency.

6.2 Intersections between planes

Once the homographies have been consistently computed as shown in the previous section, intersections between planes can be computed in a very easy and efficient way: intersection between planes $i$ and $j$ implies that a given point $x$ belonging to their intersection at the first image satisfies $H_i x \sim H_j x$; thus $x \sim H_i^{-1} H_j x$, being $x$ an eigenvector of $M_{ij}$. The eigenvectors of $M_{ij}$ are:

- the epipole $e$, which is common for all the possible couples of planes: 
  $$(I + e s_{ij}^T) e = (1 + s_{ij} \cdot e) e$$
- every vector $a$ orthogonal to $s_{ij}$: 
  $$(I + e s_{ij}^T) a = a + (s_{ij} \cdot a) e = a$$

The intersection between planes $i$ and $j$ is the set of all such vectors $a$. Furthermore, by orthogonality we know that $s_{ij} \cdot a = 0$, then the intersection (line) equation is straight given by the vector $s_{ij}$.

Our approach does not require the explicit computation of the eigenvectors of matrix $M_{ij}$, which may be unstable for several cases. Using the parameters obtained from the minimization of Equation (14):

- the intersection between the reference plane and a given plane $i$, observed at the first image, is directly given by $s_{ref,i} \cdot x = 0$;
• the intersection between any other planes \( i \) and \( j \) are given by the eigenvectors of \( \mathbf{M}_{ij} \):

\[
\mathbf{M}_{ij} = \mathbf{M}_{ref,i}^{-1} \mathbf{M}_{ref,j} = \left( \mathbf{I} + \mathbf{e} \mathbf{s}_{ref,i}^T \right)^{-1} \left( \mathbf{I} + \mathbf{e} \mathbf{s}_{ref,j}^T \right) = \mathbf{I} + \frac{\mathbf{e} \left( \mathbf{s}_{ref,j} - \mathbf{s}_{ref,i} \right)^T}{1 + \mathbf{s}_{ref,i} \cdot \mathbf{e}}
\]

Thus, the intersection line is defined by Equation \((\mathbf{s}_{ref,j} - \mathbf{s}_{ref,i}) \cdot \mathbf{x} = 0\).

6.3 Improving border detection

In order to get a better segmentation for scenes containing polyhedral structures, we also propose an improvement to the original QMMF method. The approach consists of avoiding probability propagation through the plane intersections and delaying the propagation if there is not enough information to decide to which region a pixel belongs. We reach this goal by changing the spatial coherence term in Equations (10) and (11) to:

\[
\lambda \sum_{(x,y)} z(x, y) \|p(x) - p(y)\|^2
\]

where \((x, y)\) represents two neighbor pixels \( x \) and \( y \) (left, right, up and down neighbors), and \( z(x, y) \) is a function controlling the probability propagation between these two pixel positions.

The approach is divided into two stages:

(1) For every iteration at the segmentation step, \( z(x, y) \) is computed by \( z(x, y) = \hat{\mathbf{v}}(x) \cdot \hat{\mathbf{v}}(y) \), where \( \hat{\mathbf{v}}^T(x) = (\hat{v}_1(x), \ldots, \hat{v}_N(x)) \) is the vector of normalized likelihoods at pixel \( x \).

- Normalization means that \( \sum_{k=1}^N \hat{v}_k(x) = 1 \), thus the maximum value for \( z(x, y) \) is reached only when \( \max\{\hat{v}_k(x)\} = \max\{\hat{v}_k(y)\} = 1 \); it means that the propagation will be bigger when both likelihoods associated to the same segmentation model are close to 1.
- If two or more segmentation models have similar likelihoods for neighbor pixels \( x \) and \( y \), \( z(x, y) \) will reduce the propagation between them. For example, if \( \hat{v}_k^T(x) \approx \hat{v}_k^T(y) \approx \left( \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0 \right) \), then \( z(x, y) \approx 2(\frac{1}{2})^2 = \frac{1}{2} \).
- If likelihood vectors indicate that pixel \( x \) is assigned to a given model and pixel \( y \) is assigned to a different one, \( z(x, y) \) will approximate to zero, retarding the propagation between two different regions.

(2) At a final stage, we use the border information as given in Section 6.2 in order to refine the segmentation obtained at the iterative process. We assign a very small value \( \epsilon \) to \( z(x, y) \) when the pixels \( x \) and \( y \) are separated by the intersection of planes \( i \) and \( j \), only if \( x \) and \( y \) were assigned to
one of these planes (both assigned to the same plane, or \( x \) assigned to a plane, and \( y \) assigned to the other one).

### 7 Experimental results

In this section, we present the results of segmenting stereo-pair images with the proposed method. The images present on classical benchmarks are mostly textured, and then most of the methods are not allowed to cope with problems like segmenting: single-color surfaces, different coplanar models with similar texture, or single surfaces containing several textures or colors. In our experiments, we selected stereo pair views showing special difficulties (like (i) a single plane containing different textures or colors, or (ii) different planes having the same texture, e.g the roof) that do not appear in classical benchmarks.

#### 7.1 Results on the Middlebury database

Firstly, we tested our approach on a set of three stereo pairs provided at the Middlebury dataset [16] (available online at http://www.middlebury.edu/stereo/).

For these three stereo pairs, the user has defined manually some sample regions (color circles at Figures 3-a) lying in some planar surfaces. The interface is very easy: the user click on a given point and a circular region is drawn around the point with a fixed diameter (25 pixels in these experiments). The standard deviation used to compute the registering likelihood (Equation 6) is fixed to \( \sigma = 10 \), and the binary QMMF parameters are \( \lambda = 10 \) and \( \mu = 4 \).

The stereo pairs show several slanted planes with varying amounts of texture (even with no texture as in some regions of the venus scene). For these three pairs, we computed the input likelihoods using uniquely coplanarity information (no color information has been added). The results obtained after 10 iterations of our approach are shown in Figure 3-c. The regions grew from the user-defined samples in only ten registering-segmentation steps, giving results almost ten times faster than the Lin’s method [9]. Of course, region borders are better defined for the Lin’s technique, but our method has another advantage: in the case of the cheerios and the venus scenes, Lin’s method is not capable to distinguish that the pink and green regions corresponds to different planes (cheerios), as blue and yellow regions (venus) are also different. The same problem has been found in the best methods reported in [16] for both images. Our method offers the advantage to indicate the procedure that these regions are different, and it statistically controls the region growing taken into account this fact and local coplanarity properties, giving correct region
Fig. 3. Stereo images of the Middlebury database: *cheerios* scene (first column), *sawtooth* scene (second column) and *venus* scene (third column). (a) Right images of the stereo pairs and (b) user-defined sample regions at the left images (circles with diameter = 25). (c) Resulting segmentations (after ten registering-segmentation iterations of our method for each planar model). (d) Results for the automatic Lin’s method. (e) Ground-truth images.

Another advantage of the interactivity is the ability to control the spatial coherence of the regions. At the *venus* sequence, we see that the segmented planes are hollow regions, mainly in image zones where there is no texture. In
order to force that the pixels be assigned to a given planar model, we changed the segmentation method from binary QMMF to several-models QMMF (Figure 4) and we observed the results at different coherence values. Observed gaps are due to occlusions or not connected regions. In these results, we can see the advantage of an interactive segmentation in order to easily choose the best method or parameters set for a given stereo pair; furthermore, in this step, it is not necessary to perform again the registration step, because the QMMF step is only related to the segmentation procedure.

Finally, we tested two different registering strategies: (i) considering that planar homographies are completely independent, (ii) taken into account homography dependency by using the projective framework presented in section 6. For the first registering approach, tests require several hundred of iterations, taking several hours before to converge to a stationary solution. On the other hand, considering consistency allows to converge in very few iterations, as it is shown in Figures 3 and 4. When convergence is reached, the solutions for both approaches are almost the same, such that differences are difficult to visualize; hence, the main advantage on considering homography consistency is on improving the computational time.

7.2 Refining the plane-color image segmentation

In Figure 5-a, we appreciate that the obtained regions seem to be qualitatively similar to the ones stated in the ground truth. Nevertheless, the borders of the regions are not well-defined. In order to improve the final segmentation, we add the refining step that takes into account the intersection between planes as it is indicated in Section 6.3. Figure 5-b shows the lines corresponding to plane intersections that are visible in the first image. At Figure 5-c, it is shown that the region growing is controlled in order to avoid propagation through the regions borders and it helps to improve the segmentation quality for polyhedral scenes.
Fig. 5. Segmentation of the cheerios scene, (a) for ten iterations of our approach, (b) the map of plane intersections, (c) the image segmentation taking into account plane intersections.

7.3 Including color and texture to planar models

The tests were also conducted on another set of stereo views of real outdoor scenes. The experiment shown in Figure 6 is designed to demonstrate the importance of incorporating context information related with texture in the likelihood computation. Experiment shown in Figure 7 illustrates the importance of taking into account both color and coplanarity for computing likelihoods. In this section we intend to show step by step how informative are coplanarity, texture and color likelihoods as inputs for the segmentation process.

Fig. 6. Normalized coplanarity likelihoods computed (a) pixel-to-pixel and (b) for $3 \times 3$ windows, for both planes of the tower, left and right.

Firstly, we compute the homographies corresponding to both observed planes at the tower scene by means of the procedure described at section 3. The
Fig. 7. Likelihoods computed for (a) planar similarity and (b) coplanarity and color.

Normalized likelihoods obtained for these registering models, considering pixel-to-pixel similarities, are shown in Figure 6-a. White pixels represent highest similarities, while darker pixels correspond to lower planarity fitting to that model. Most of the pixels corresponding to the correct plane are shown in white; nevertheless, the regions are too granular and not connected. When using $3 \times 3$ regions (as it is shown in Figure 6-b), the coplanarity likelihoods are improved and well-defined, such that they can easily be distinguished by very simple methods, like thresholding the images.

Nevertheless, normalized planarity likelihoods may be less informative if the regions are untextured or if there are different coplanar models with similar texture. This is the case of the left and right sides of the roof (shown in Figure 1). In the images at Figure 7-a, only the pixels corresponding to high gradient points at the original images have high likelihood. Flat-color regions have almost the same likelihood for the different models, as it is observed for the sky pixels and the facade. Moreover, intersections between planes are not well-defined, because the border regions may correspond to either plane at the junction. When color information is included, highest likelihoods describe better the complete planar surfaces and not only the high gradient elements, as it is shown in Figure 7-b.

7.4 Results on outdoor images

At this step, maximum likelihood estimator (MLE) is good enough to distinguish most of the coplanar pixels with similar texture and color properties, but the detected regions are granular and not necessarily connected. In order to solve this problem, we apply the QMMF segmentation method (described in section 5). Once the a posteriori probabilities were obtained, we select the planar regions by selecting the pixels whose probability is greater or equal than 0.5 and, then, the biggest connected region for each planar model.

If the results are not satisfactory at the end of the complete registration-
Fig. 8. Experiments for the: tower scene (first column), roof scene (second column) and house scene (third column). (a) Small user-defined sample regions at the left images and (b) their corresponding segmentations (after 10 binary QMMF iterations); (c) large user-defined sample regions and (d) the resulting segmentations (after just one binary QMMF iteration for each model). Note the segmented regions correspond to class membership larger than 0.5.

For the final experiment, the user defined some image samples, shown in Figure 8-a, and the obtained results after ten binary QMMF iterations are shown in Figure 8-b. In the three sets of images, the planar regions were correctly extended, but there are still misclassifications if there are two concurrent planes or regions with similar texture or color. In Figure 8-c, the user extends the sample regions such that the results shown in Figure 8-d were obtained after
only one binary QMMF iteration. Our results also show that the algorithm can distinguish planar surfaces from non-planar elements (e.g. in the ‘roof’ stereo pair, the tree leaves are correctly segmented out from the left region of the house roof).

8 Summary and conclusions

In this paper, we have proposed a method for segmenting and registering coplanar regions of a scene, based on the observation of a pair of stereo views. This process can be used in any environment containing planar structures. Its domain of application is wide because planar surfaces are quite common both at indoor and outdoor scenes. One can intrinsically segment a piecewise planar scene from two 2D images without performing neither camera calibration nor 3D reconstruction.

There are still limitations and further research to continue. We have chosen an assisted strategy to initialize our algorithm. Indeed, this approach depends on the user selection of sample regions and it may be sensitive to this choice. Although automatic planar segmentation methods have been found in the literature, interactivity may help to reduce computational time (from several hours [9] to some seconds in our method) and to correct wrong segmentations. However, we consider that this stage is independent of our method and that it does not demerit our proposal; contrarily, it may help to extend planar regions that other methods can not detect or separate. Our approach may accept any other initialization method and we are in actuality studying other approaches for this step.

We also have analyzed the information that can be extracted from color properties of the images and from move matching between two views of a planar surface. We propose to use an approach combining both kinds of information and to refine the marginal probability of finding both criteria for a given planar model, by means of an innovative Bayesian approach based on Markov random fields.

One of the future improvements of our method is that now registering has been conducted only from the first image to the second one, but no coherence from the registering in the opposite sense has been verified. This step should improve the performance of our method in some challenging regions such as discontinuities and occlusions.

Another goal of our future work consists on obtaining metric information from the computed registering parameters, such as the epipole positions and the plane parameters, for a larger set of views, intending to auto-calibrate the
cameras and to recover the planar piecewise structure of the scene.

References


