A PATCH RECTIFICATION STRATEGY FOR MULTI-HOMOGRAPHY ESTIMATION

Mario Santes

Pascal Institute, UMR6602 Blaise Pascal University, CNRS, Campus des Cézeaux, 24, Avenue des Landais. 63177 Aubiere, France.
ms@sant3s.com

Keywords: Polyhedral Scene, Multi-planar Scene, Homography, Epipolar Geometry.

Abstract: In this work, we describe how to recover the perspective functions induced by the dominant faces of a rigid polyhedral scene captured by two cameras at different positions. The faces of any polyhedron can be dealt with as if they were planar surfaces, therefore we analyze this problem in a multi-planar scene context. We assume that the scene is populated with polyhedral objects and that a regular grid is imposed on the reference image. We describe a robust strategy that is able to merge patches that belong to the same planar surface and to impose the epipolar geometry constraints to the recovered perspective functions.

1 INTRODUCTION

In man-made environments, the presence of polyhedral is common. These surfaces are present as walls, doors and many objects such as books, desks and other furniture. The Computer Vision community has been working very hard in order to describe this special kind of structures and to recover the camera motion when dealing with two or more views of the same rigid scene. The faces of any polyhedron can be dealt with as if they were planar surfaces, therefore we can analyze that environment in a multi-planar rigid scene context. An example of polyhedral scenes can be observed at Fig. 1.

Figure 1: Images projecting two different scenes. The left hand side image has only one dominant surface. The second view has two dominant planar surfaces. In both images, the grass is not useful, due to the homogeneous texture.

Two different views projecting the same scene are enough in order to recover camera motion and the scene’s structure (Hartley and Zisserman, 2004).

Structure recovery can be dealt with in at least two ways for this kind of scenes: (i) by reconstructing the point cloud that belongs to the observed surface (Kanatani and Niitsuma, 2010), or (ii) by recovering all the surface’s parameters by using two views and the inter-image transfer function induced by the planar surface (Malis and Vargas, 2007). In any of the cases given above, it is assumed that all the points belong to a planar surface. On the other hand, RANSAC-like robust estimators can be used in order to fit a planar model to the reconstructed point cloud or directly on the image domain, by fitting the inter-image transfer function induced by the planar surface. In this work, we show how to recover the dominant inter-image transfer functions based on a patch rectification strategy on the image domain. The patch-rectification approach presented in this work uses the following conjecture: we assume that a regular lattice is superimposed on the first image. Lattice elements or patches, capture small regions of the observed scene, see Fig. 2. For each patch, we compute the inter-image function that transfers points from the first image to the second image. Under this mapping, reconstruction parameters can be obtained, i.e. if the patch is generated by only one planar surface, the inter-image function can be decomposed to recover the surface’s parameters and the rotation and translation parameters for the second camera, by assuming that the first camera is located at the reference frame. When two or more planar surfaces are captured by
the same patch, the induced mapping does not allow to recover useful information.

![Image of two views projecting a polyhedral scene](image)

**Figure 2:** In this figure, we sketch two views that project a polyhedral scene, consisting of three faces or bounded planar surfaces. A grid has been superimposed on the first view. Each view contains many feature points. The intersection of any two planar surfaces is represented by a line segment or straight line.

## 2 RELATED WORK

Image’s sparse segmentation using algebraic and geometrical constraints was studied in (Santes and Vigueras, 2009). That work presents a set of linear equations that allows to incorporate the geometric and algebraic constraints during the computation of inter-image transfer functions, the first epipole and the projections of the lines representing the intersection of the planar surfaces. In (Vigueras and Rivera, 2010) a user-supervised methodology is proposed in order to obtain a maximum likelihood estimation of the inter-image transfer functions with epipolar geometry consistency without explicit computation of the fundamental matrix. Homography estimation for one planar surface is a well understood task, (Hartley and Zisserman, 2004). However, it is not well understood how to merge the support pixels from any two or more homographies when those homographies were induced by the same planar surface. This problem occurs when we assume that the image’s generative model consists of a collection of non-overlapping patches that project only a planar surface. A naive approach to deal with this problem may consist of any of the following steps: (i) computing the norm of the difference between any two homographies, (ii) decomposing the homographies in order to recover the plane normals.

### Problem definition and contributions

In this work, we assume that we have two images of the same rigid polyhedral scene with two or more dominant planar surfaces. We want to recover the homographies induced by the dominant planes. A patch rectification methodology is proposed in order to recover the homographies associated to the dominant planes. This technique is based on non-overlapping patches that are rectified via the solution of the multi-homography equation. In this work, feature points are used to compute the dominant homographies.

## 3 PROPOSED APPROACH

We assume the pin-hole camera model with non-normalized coordinates. A planar surface is parameterized by a normal vector $n \in \mathbb{R}^3$ and a scalar $d$ such that if $X \in \mathbb{R}^3$ is a space point that lies on the surface, then $n^T X + d = 0$. The inter-image function induced by observing a planar surface is called an homography $H \in \mathbb{R}^{3 \times 3}$ and it is defined as:

$$H = K_2 \left( R - t_n n^T \right) K_1^{-1},$$

where $K_1, K_2 \in \mathbb{R}^{3 \times 3}$ are the camera matrices for the first and second views, respectively. $R \in SO(3)$ and $t \in \mathbb{R}^3$ are the rotation and translation parameters. Consider $x$ and $x'$ as image points on the first and second view respectively. If these points are the projection of a 3D point lying on a planar surface that induces the homography $H$, then we write this relation as: $x' \sim Hx$. The symbol $\sim$ means that there exists a non-null scalar such that the above relation becomes an equation. The $l-th$ correspondence pair is denoted as: $<x \leftrightarrow x'>$. Image points $x$ and $x'$ are represented in homogeneous coordinates. Define the planar homology $M_{i,j}$ as follows: $M_{i,j} = H_j^{-1} H_i$. By using $M_{i,j}$ and the Sherman-Morrison formula, we obtain:

$$M_{i,j} = H_j^{-1} H_j = I + es_{i,j}^T,$$

where: $e = K_i R^{-1} t$, $s_{i,j} = K_i^{-1} (\dot{v}_i - \dot{v}_j)$ and $\dot{v}_i = \alpha_i n_i$ for some $\alpha_i \in \mathbb{R} \setminus \{0\}$, the same applies for the $\dot{v}_j$ vector. The $s_{i,j}$ vector is the projection at the first image of the intersection between planes $i$ and $j$, it follows that $s_{i,j} = 0, \forall i$. $e$ is the first epipole. Given the planar homology $M_{i,j}$, we represent $H_j$ as follows:

$$H_j = H_i \left( I + es_{i,j}^T \right).$$

(1)

Once $H_i$ and the $\{s_{i,j}\}$, $e$ vectors have been estimated, remaining homographies can be expressed by using Eq. (1). In order to solve Eq. (1), a first approximation for the epipole $e$ can be obtained by decomposing the following equation: $M_{i,j} - I = es_{i,j}^T$, by means of the SVD algorithm. After the epipole has been estimated, the $s_{i,j}$ vector can be obtained by solving the linear equation: $H_j^{-1} x' \sim Ix + es_{i,j}^T x$. Define the following matrices: $D = [H_j^{-1} x'] x e^T \in \mathbb{R}^{3 \times 3}$ and $b = -[H_j^{-1} x'] x \in \mathbb{R}^3$. The support set $H_j = <
\( x_j' \leftrightarrow x_j'' > \) for homography \( \mathbf{H}_j \) must be used to generate a linear system as follows:

\[
\begin{pmatrix}
\mathbf{D}'^1 & \vdots & \mathbf{D}'^m
\end{pmatrix}
\begin{pmatrix}
\mathbf{b}'^1 \\
\vdots \\
\mathbf{b}'^m
\end{pmatrix}
= \begin{pmatrix}
\mathbf{s}_{i,j} \\
\vdots \\
\mathbf{s}_{n,j}
\end{pmatrix},
\]

where \( \mathbf{D}' \) is the matrix corresponding to the \( l \)-th feature pair inside the support of the homography \( \mathbf{H}_j \), considering \( \mathbf{H}_i \) as the reference homography. \( m_j \) represents the size of the support of the homography \( \mathbf{H}_j \), i.e. \( m_j = |\mathbf{H}_j'|. An improved estimation of the first epipole \( \mathbf{e} \) can be calculated by a similar procedure, therefore: \( \mathbf{H}_j^{-1}x' = (x's_{i,j})e = -\mathbf{H}_j^{-1}x'x \).

As before, we set \( \mathbf{D} = [\mathbf{H}_j^{-1}x' | (x's_{j,i})] \in \mathbb{R}^{3 \times 3} \) and \( \mathbf{b} = -\mathbf{H}_j^{-1}x'x \in \mathbb{R}^3 \). The support of each homography must be used to improve the estimation of the epipole, by doing this we obtain the following system:

\[
\begin{pmatrix}
\mathbf{D}_1^1 \\
\vdots \\
\mathbf{D}_1^m
\end{pmatrix}
\begin{pmatrix}
\mathbf{b}_1^1 \\
\vdots \\
\mathbf{b}_1^m
\end{pmatrix}
= \begin{pmatrix}
\mathbf{e} \\
\vdots \\
\mathbf{e}
\end{pmatrix},
\]

where \( \mathbf{D}_1 \) is the matrix corresponding to the \( l \)-th feature pair for the \( j \)-th plane. These linear systems are over-determined. Suppose that \( \mathbf{H}_i \) represents a reference homography. \( \mathbf{H}_j \) can be re-estimated by including the support of all remaining homographies in conjunction with its own support. Given that \( x'^j \sim \mathbf{H}_j(I + \mathbf{e}s_{i,j})x' \), we have: \( \mathbf{y}' = (I + \mathbf{e}s_{i,j})x' \). Therefore, \( \mathbf{H}_j \) is re-estimated from the following relation: \( x'^j \sim \mathbf{H}_j \mathbf{y}' \).

Non-linear adjustment. In order to improve the previous estimation, we propose to solve the non-linear problem:

\[
\min_{\{\mathbf{H}_i, \mathbf{e}, \mathbf{s}_{i,j}\}} \sum_{j=1}^n \sum_{l=1}^{m_j} \| x_j' - \zeta \left( \mathbf{H}_i (I + \mathbf{e}s_{i,j}) x_j' \right) \|^2,
\]

where \( \zeta \) is a normalization function, defined as: \( \zeta(x, y, z) = (\frac{x}{z}, \frac{y}{z}, 1)' \). This non-linear problem, can be solved by means of the Levenberg-Marquardt Algorithm \( \text{(Noisedal and Wright, 1999)} \). We use the previous linear systems, Eqs. (2) and (3), to find a suboptimal solution that allows us to reach a close-to-optimum solution avoiding local minima.

**Patch rectification.** Let \( I_1 \) be the first image, and \( x \leftrightarrow x' \) the set of correspondence pairs. \( I_1 \) is divided in non-overlapping regular patches, see Figs. 3 and 4. For each patch, an homography is computed by using the parameter-free LMedS Algorithm with normalization of coordinates \( \text{(Hartley and Zisserman, 2004)} \). We obtain a set of homographies \( \{\mathbf{H}_j\} \). Without loss of generality, set \( \mathbf{H}_1 \) as the reference homography. Given that the homographies were computed in an independent way, they do not reflect the same epipolar geometry; by solving Eq. (1), we can express all the remaining homographies as \( \mathbf{H}_j = \mathbf{H}_1(I + \mathbf{e}s_{i,j}) \). It is not practical to compare the rectified homographies with the reference homography. It is evident that if \( \mathbf{H}_j \) shares the same planar surface’s parameters as \( \mathbf{H}_1 \), then \( \mathbf{s}_{1,j} = \mathbf{0} \).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{patch.png}
\caption{Two contiguous patches. In this figure, we sketch two patches and the feature points inside them. The line \( s_{1,2} \) lies over patch 2. This line represents the projection on the first view of the intersection between two planes. In this example, the \( s_{1,2} \) vector induces a partition over the feature points. See text and Fig. 4.}
\end{figure}

The \( \{\mathbf{s}_{i,j}\} \) vectors play an important role in our framework, they define when the patch \( j \) must be merged to the reference patch. If patch \( j \) projects the same planar surface than the reference patch, the support of both patches must be used to improve the estimation of the reference homography. This means that the patch \( j \) vanishes and therefore \( \mathbf{s}_{1,j} = \mathbf{0} \).

In the presence of noise, we must take into account two factors to determine the goodness of an homography: (i) the size of the support set \( \mathbf{H}_j \), and (ii) the average error \( \mathbf{H}_{\text{error}} \) computed from the support set. These two factors are very effective to define the weight of the inter-image functions. We propose the following weight function:

\[
\omega(\mathbf{H}) = |\mathbf{H}'| \exp(-\mathbf{H}_{\text{error}})
\]

where \( |\cdot| \) stands for the size of the support set. We define \( \mathbf{H}_{\text{error}} \) as follows:

\[
\mathbf{H}_{\text{error}} = \frac{1}{|\mathbf{H}'|} \sum_{j=1}^n d(\zeta(\mathbf{H}_j \mathbf{x}', \mathbf{x}')^2 + d(\mathbf{x}', \zeta(\mathbf{H}_j^{-1} \mathbf{x}'))^2),
\]

where \( d(\cdot) \) is the euclidean distance function. After applying the weight function, \( \omega \), we obtain a set of ranked homographies. The patch rectification procedure explained above, must be applied considering the ranked set of homographies. Algorithm 1 summarizes the theory described in this section.
Figure 4: Patch rectification example. Images are described from left to right. (1) The grid superimposed on the first view with the reference patch in red. The lines represent the \( s_{1,j} \) vectors. (2) Patches selected to be merged to the reference patch, and (3) the feature points belonging to the projection of the same planar surface. The \( \{ s_{1,j} \} \) vectors induce a partition on the image that may be useful to classify feature points. However, this criteria is not enough in all the cases, see the experiments at the end of this work.

**Algorithm 1** Patch rectification

**Input:** \( <x^l \leftrightarrow x^l> \).

**Output:** \( H_1 \) and support set \( H'_1 \) for \( H_1 \).

1. Compute a regular non-overlapping patch division on the image coordinates.
2. For each patch, determine the points \( \{ x_i \} \) that lie inside this patch.
3. Compute the set \( \{ H_j \} \).
4. Apply the rank function, Eq. (5), to each element of \( \{ H_j \} \).
5. Pick the first element in the ranked set \( \{ H_j \} \), set this element as \( H_1 \).
6. Solve Eq. (1) by using Eqs. (2) and (3). In order to obtain a first guess for the epipole, solve \( M_{1,j} \) for all \( j \). Repeat until convergence.
7. Solve the optimization problem, Eq. (4), if needed.
8. For each patch \( j \), if \( s_{1,j} \rightarrow 0 \), merge patch \( j \) with the reference patch.
9. Re-estimate \( H_1 \) with the LMedS Algorithm.
10. Re-label the new inliers obtained in the previous step, as elements of the reference patch, i.e. update \( H'_1 \).

4 Experiments

To apply our patch rectification approach, we require two images of the same polyhedral scene. We create a lattice on the first image by considering regular patches. We use SIFT descriptors [Lowe, 2004] in order to obtain the set \( <x \leftrightarrow x> \). However, our approach only requires points inside the image domain.

**Synthetic data.** We conduct experiments over a synthetic scene, consisting of a cube with many levels of gaussian noise with standard deviation \( \sigma \in \{0.5, 1.0, \ldots, 6.0\} \) pixels. The additive noise was applied to the correspondence pairs \( <x \leftrightarrow x> \). The regular mesh used in these experiments contains \( 10 \times 10 \) patches. This grid includes patches that project two or three of the simulated polyhedron’s faces. We compute three parameters to evaluate the quality of the fitted parameters: (i) the Frobenius norm of the difference matrix between estimated homographies and ideal homographies, Fig. (5b), (ii) the 2-norm of the difference of the intersection between planes, Fig. (5c), (iii) the angle between the estimated and ground-truth values of the intersection between planes, \( \{ s_{1,j} \} \) vectors, Fig. (5d). With this setting, we can see that our approach is robust to high levels of gaussian noise. Figs. (5b), (5e) show that even with high noise levels (i.e. \( \sigma \leq 4.0 \)) the \( \{ s_{1,j} \} \) vectors do not present high variations w.r.t. the ground-truth parameters. To validate our approach w.r.t. the epipolar geometry, we recover the fundamental matrix \( F \) by using (i) the reference homography \( H_1 \), and also (ii) all the correspondence pairs that belong to the detected planar surfaces, see [Hartley and Zisserman, 2004]. \( F \) is determined from \( H_1 \) by using \( F = [e] \cdot H_1 \), with \( e = H_1 e \). Fig. (5f) shows that our proposed approach outperforms the direct method.

**Real data.** In this section, we present four experiments with indoor and outdoor environments. The grid’s granularity strongly depends on the image’s size and on the distribution of the feature points. However, increasing the granularity of the superimposed grid, increases the number of independent homographies to compute. See Fig. 4 (a) Maps sequence. Our first experiment consists of a perfect polyhedron with two visible faces. The proposed algorithm was able to detect two homographies \( H_1 \) and \( H_2 \). In this case, the line \( s_{1,2} \) is part of the scene. In order to exemplify how our proposed approach works, we draw the grid that was superimposed on the first view. As it can be observed, the \( s_{1,2} \) vector induces a partition on the first image. This fact can be useful in this kind of scenes, in order to classify or segment the correspondence pairs. (b) Desktop sequence. This is another experiment with two pla-
The flat screen induces the dominant homography. Detected feature points that lie at the same level that the keyboard generate a planar surface that induce the second homography, with an average error a bit high, when comparing to the reference homography. In this example, the line $s_{1,2}$ is not part of the scene, although it can be deduced from the scene’s geometry. (c) Wadham sequence. This experiment, as opposed to the previous one, shows how the $s_{1,2}$ vector can not be used as a segmentation criteria for the correspondence pairs set. In this case, the $s_{1,2}$ line is visible but a small planar surface, at the roof, is coplanar to the reference plane. (d) Corridor sequence. The images used for this experiment are a bit challenging, due to the lack of texture. In our work, we use only feature points. It’s well known that a matching algorithm based on lines gives better results for this dataset, (Hartley and Zisserman, 2004). We recover three homographies, the first and third detected planar surface have a very low average error (less than 1 square pixel). The second homography has some outliers inside its support set. Wadham and Corridor datasets are available at http://www.robots.ox.ac.uk/~vgg/data

5 CONCLUSIONS

The proposed approach presented in this report, allows to compute robust inter-image homographies from polyhedral scenes with minimal information and without specifying precise thresholds for random estimators. Our patch-based strategy is able to take one patch and to impose same motion parameters to remaining patches. After applying this strategy, the inter-image functions codify the same epipolar geometry, without explicit computing of this geometry. Our approach is suitable for initializing tracking algorithms for polyhedral scenes and we strongly believe it can benefit visual servo control algorithms based on polyhedral in real environments. On the other hand, image segmentation algorithms with multi-planar constraints (c.f. (Vigueras and Rivera, 2010)) can avoid the parameter estimation stage by using our proposed approach. However, multi-planar dense image segmentation with no well-textured regions is still an open problem.

Acknowledgements. This work was supported by The Auvergne Region, France.

REFERENCES


Figure 6: (a) Maps sequence. This experiment was carried out with two views of a map. The image’s size is 640 × 480 pixels. In order to exemplify our proposed approach, we use a low granularity grid, consisting of only 4 × 4 regular patches. The third image corresponds to the grid computed on the first image. The marked patches lie on the intersection between the two planar surfaces. The $s_{1,2}$ vector can be used to segment the final support sets. The last image corresponds to another viewpoint of the same scene. (b) Desktop sequence. Images’ size is 640 × 480 pixels, the grid on the first view has 8 × 8 regular patches. The flat screen acts as the reference plane, with 785 correspondence pairs and an average error of 2.39834 pixels$^2$. The second planar surface has 227 correspondence pairs with an average error of 6.62609 pixels$^2$. (c) Wadham sequence. Images’ size is 1024 × 768 pixels, the grid on the first view has 10 × 10 regular patches. In this sequence, the grass was not successfully detected by our implementation. The third image shows the support for the recovered homographies and the $s_{1,2}$ vector. Reference plane is at the right side. $H_1$ has 1388 feature points with an average error of 1.85394 pixels$^2$. $H_2$ was computed with 891 correspondence pairs with an average error of 1.43526 pixels$^2$. This example shows how our approach allows to recover small regions, see the $H_1$’s support set. (d) Corridor sequence. This sequence is quite challenging due to the lack of texture. The image’s size is 512 × 512 pixels, the regular mesh has 8 × 8 patches. The reference plane is at the corridor’s floor (red). The computed average error for $H_1$ is 0.460016 pixels$^2$, the support set has 309 elements. The second planar surface (green) is at the top-center of the image. The associated homography has an average error of 1.26612 pixels$^2$ computed with 209 pairs. This homography still includes many outliers. The third planar surface is at the column (blue), the homography has an average error of 0.976872 pixels$^2$ with 195 correspondence pairs.