Eur. Conf. on Comp. Vis., Springler Verlag, ECCV'00, Dublin, june 2000

Geometric driven optical flow estimation and segmentation for 3D reconstruction

Lionel Oisel, Étienne Mémin, and Luce Morin

IRISA, Campus de Beaulieu, Rennes, France

Abstract. We present a method for fully automatic 3D reconstruction from a pair of uncalibrated images in order to deal with the modeling of complex rigid scenes. A 2D triangular mesh model of the scene is calculated using a two-step algorithm mixing sparse matching and dense motion estimation approaches. The 2D mesh is iteratively refined to fit any arbitrary 3D surface. At convergence, each triangular patch corresponds to the projection of a 3D plane. The algorithm proposed here relies first on a dense disparity field. The dense field estimation modelized within a robust framework is constrained by the epipolar geometry. The resulting field is then segmented according to homographic models using iterative Delaunay triangulation. In association with a simplified self-calibration algorithm, this 2D planar model is used to obtain a VRML-compatible 3D model of the scene.

Many recent works attempt to deal with 3D reconstruction from set of images. Two different classes of approaches are generally proposed using different types of information: the first one includes model-based methods and the second one deals with model-free methods.

In model-based approaches, the scene information is assumed to be composed of large polygonal objects described by a limited set of 3D points characterizing the vertices of each 3D plane. This model can be computed in the 2D space without 3D information. This can be done by extracting, matching and 3D reconstructing points of interest[6] or edges [7]. One of the main limitations of these methods is that effective planarity of generated facets is assumed but not always satisfied. To enforce a global planarity, a manual intervention is even usually necessary to indicate reliable coplanar points. Another way of estimating the model is to use disparity maps (or alternatively depth maps). In [9], Koch *et al.* suggest computing differential properties from a dense disparity map. Images are then segmented according to similar surface orientation at each point of a region. The underlying strongly polyhedral assumption is indeed the major limitation of model-based techniques.

To enlarge the variety of treated scenes, model-free representations (second class of approaches) have been proposed. Such methods generally rely on a dense disparity map. This map can be combined with weak or strong calibration information to provide a depth map that can be manipulated for view synthesis [6, 8]. The major limitation consists here in the estimation of reliable dense disparity information allowing occlusion areas and spatial discontinuities to be coped with efficiently.

The main objective of our study is to propose an entirely automatic approach for the reconstruction of not necessarily polyhedral textured scenes. In addition to this non-specialized goal, we impose to have the ability of an easy and real time visualization. This latter requirement dismisses practically the use of methods based entirely on a dense depth map. On the other hand, the removal of the polyhedral scenes assumption favors such approaches. Following these two remarks and in order to comply with the previously described goals, our aim is to suggest a compromise between model-free and model-based methods. We first propose to describe the 3D scene by a triangular mesh which can be displayed by most visualization dedicated systems. Our method therefore belongs to the first class (model-based approaches) but as this triangular mesh is automatically computed from a dense disparity field, it is also related to the second class.

The key point of our method is to segment the images into regions which are actually planar in the 3D scene and to extract the planarity propriety from the image data (and not from a user intervention). This is indeed equivalent to realizing motion segmentation according to an homographic model. As the homographic model describing the set of admissible transformations of planar patches is non-linear, a direct region-based segmentation method is hardly feasible. We have therefore designed a two-step method. The first step provides a geometrically constrained dense depth map and an associated discontinuity map. This dense information is then used to initialize the second step: homographic model estimation and segmentation.

The outline of the paper is the following. The first section briefly describes geometric definitions associated with perspective projection of two images. In this context, epipolar geometry is presented. This important geometric constraint is used in all the following steps of our method and has to be previously estimated.

In the second section, in order to facilitate a subsequent planar facet segmentation step, we present a geometrically constrained disparity field estimation. This technique is derived from a robust optical flow estimation approach. Unlike classical correlation methods, it provides a reliable piecewise smooth motion field [2, 12]. Moreover the disparity estimation is constrained by the associated epipolar geometry so that the estimated field is explicitly forced to be geometrically consistent with a perspective projection model and with the fixed scene assumption. This constraint also yields a substantial computational cost decrease (the 2D disparity estimation problem is reduced to a 1D problem).

The third section presents the planar facet segmentation step of our method. To ensure the effective planarity of each reconstructed triangle, an adaptive iterative triangulation based on homographic models estimation is computed from the disparity field.

By arbitrarily fixing intrinsic parameters, 3D rotation and translation parameters can be extracted from the epipolar geometry. Using this 3D information, the resulting 2D model is then re-projected in the 3D space to be visualized as a VRML representation.

This method has been validated on synthetic and real world images. Comparison with existing classical techniques are presented in the last section of the paper.
Remark: in the following, vectors will be represented by bold letters.

1 Epipolar geometry

1.1 Definition

The characterization of the geometry associated with the two cameras is of key importance in order to build a 3D model of the scene. In our case, we deal with two uncalibrated cameras (or alternatively one moving camera shooting a rigid scene) assuming a pinhole camera model. This model characterizes the projection of a 3D point $\mathbf{P}(X,Y,Z)$ on a point $\mathbf{p}(x,y)$ of the image plane. In the case of two images, the projection model is defined by a system of two equations linking a 3D point $\mathbf{P}(X,Y,Z)$ to its projections $\mathbf{p_1}(x,y)$ in the first image and $\mathbf{p_2}(x',y')$ in the second one. Without lost of generality, we assume that the world coordinate system coincides with the first camera coordinate system. The resulting system can be written using homogeneous coordinates as follows (where $\tilde{}$ denotes homogeneous coordinates):

$$\begin{aligned} \tilde{\mathbf{p}}_\mathbf{1} &= A_1[I\ 0]\tilde{\mathbf{P}} \\ \tilde{\mathbf{p}}_\mathbf{2} &= A_2[R\ \mathbf{t}]\tilde{\mathbf{P}} \end{aligned} \tag{1}$$

R is the rotation matrix and $\mathbf{t}$ the translation vector between the first and the second camera location (extrinsic parameters). Matrix A contains internal camera parameters (intrinsic parameters).

Eliminating $\tilde{\mathbf{P}}$ in equations (1) leads to a relation linking the projections of a 3D point in both images:

$$\tilde{\mathbf{p}}_\mathbf{2}^T {A_2}^{-T}[\mathbf{t}]_\times R {A_1}^{-1}\tilde{\mathbf{p}}_\mathbf{1} = 0, \tag{2}$$

where $[\mathbf{t}]_\times$ denotes the cross product matrix associated with the translation vector.

This constraint called epipolar geometry has been first introduced by Longuet-Higgins [11]. It is entirely defined by a 3×3 homogeneous matrix called the *fundamental matrix* formulated as $F_{1\,2} = {A_2}^{-T}[\mathbf{t}]_\times R {A_1}^{-1}$. By construction, this matrix is of rank 2 and is defined up to a non-zero scalar factor. A fundamental matrix has therefore only seven degrees of freedom.

The epipolar constraint can be used to determine the epipolar line $\mathbf{l_2}$ in the second image associated to the point $\mathbf{p_1}$. It represents the line of $\mathbf{p_1}$ potential correspondences in the second image. Line $\mathbf{l_2}$ is given by:

$$\tilde{\mathbf{l}}_\mathbf{2} = \mathbf{F}_{1\,2}\tilde{\mathbf{p}}_\mathbf{1} \tag{3}$$

where $\tilde{\mathbf{l}}_\mathbf{2}$ denotes homogeneous coordinates of $\mathbf{l_2}$, i.e. all points in $\mathbf{l_2}$ satisfy $\tilde{\mathbf{l}}_\mathbf{2}\tilde{\mathbf{p}}_\mathbf{2} = 0$.

1.2 Case specific application

The issue we are concerned is the recovery of 3D information from sets of 2D images. It consists in solving system (1) to obtain the 3D point $\mathbf{P}$. To that end, corresponding points $\mathbf{p_1}$ and $\mathbf{p_2}$ and calibration parameters (extrinsic and intrinsic parameters) giving A_1, A_2, R, $\mathbf{t}$ have to be recovered. The first issue can be greatly simplified by constraining the matching process with the epipolar geometry while the second one can be achieved using a decomposition of the fundamental matrix (see section 4). The epipolar geometry estimation is indeed a crucial key point of our method. The next paragraph will present the method we use to recover the fundamental matrix from two uncalibrated images.

1.3 Fundamental matrix estimation

We assumed here that corresponding points have been extracted and matched using an Harris and Stephens detector associated with a cross correlation process. This first step is equivalent to the one developed by Zhang [17].

To take into account the nullity of the fundamental matrix determinant, we followed a method proposed by Boufama *et al.* based on the virtual parallax [4]. This method may be briefly described as follow. The fundamental matrix is first estimated from 8 matches: three of them are selected to perform a projective change of basis to constraint the matrix to be of rank 2. A fourth arbitrary pair is also added to complete the projective change of basis [4]. The four last pairs are then used to provide a unique fundamental matrix solution which respects the rank 2 constraints (determinant of null value).

In association with the determinant nullity constraint, the change of basis provides a normalization effect on points coordinates: the coordinates of the three points selected to characterize the new basis are assigned to values between 0 and 1. This involves that coordinates of points belonging to the triangle defined by these points, also belong to the range of 0 to 1. In order to perform an optimal normalization, the pairs of points are chosen as near as possible to image corners.

Besides, to cope with erroneous matches, a robust estimation based on least median squares estimation is incorporated [14].

2 Dense disparity field estimation

2.1 Constrained optical flow expression

Let $I_i(s)$ be the intensity in the ith image, where $s(x, y) \in S$ denotes the spatial position on grid S. Assuming a constant intensity along motion trajectories, the brightness constancy assumption is expressed as:

$$DFD(\mathbf{s}, \mathbf{d}_s) = I_1(\mathbf{s}) - I_2(\mathbf{s} + \mathbf{d}_s) \;=\; 0 \;, \tag{4}$$

where DFD stands for the Displaced Frame Difference function and $\{\mathbf{d}_s = (d_x, d_y),\ \mathbf{s} \in S\}$ for the image displacements from position 1 to position 2. In

the general case, this is a 2D problem: for each pixel, d_x and d_y have to be recovered.

Using the epipolar constraint, it is possible to decompose the displacement vector $\mathbf{d}_s$ into normal and tangential components with respect to the epipolar line (see Figure 1). The brightness constancy assumption is therefore rewritten as: $DFD(\mathbf{s}, \mathbf{d}_s) = I_1(\mathbf{s}) - I_2(\mathbf{s} + \mathbf{n}_s + \lambda_s \mathbf{v}_s) = 0$. The normal component $\mathbf{n}_s$ and the unit vector on the epipolar line $\mathbf{v}_s$ can be computed from the fundamental matrix for any position $\mathbf{s}$ (see eq. 3). The enforcement of the epipolar constraint at every point reduces the original 2D estimation problem to a 1D problem: the estimation of $\lambda = \{\lambda_s,\ s \in S\}$ along epipolar lines.

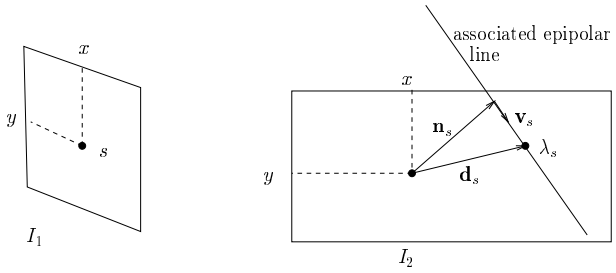

Fig. 1. *Displacement vector decomposition.*

The DFD expression is highly non linear with respect to the displacements. To avoid a tough non linear estimation, a Taylor expansion of this equation is considered around point $\mathbf{s} + \mathbf{n_s}$. This linearization leads to a constrained optical flow equation:

$$I_1(\mathbf{s}) - I_2(\mathbf{s} + \mathbf{n}_s) - \lambda_s \mathbf{v}_s \nabla I_2(\mathbf{s} + \mathbf{n}_s) = 0$$

where ∇ is the spatial gradient.

This equation relies nevertheless on an inherent ambiguity. The fundamental matrix defining epipolar lines is well known to be far more reliably estimated for large displacement between two camera view points. This large displacement assumption somewhat contradicts the infinitesimal disparity hypothesis. To overcome this incompatibility, the estimation is embedded in a coarse-to-fine multiresolution scheme.

2.2 Multiresolution scheme

At a given level k, the disparity $\lambda_s = \{\lambda_s^k,\ \mathbf{s} \in S\}$ is decomposed into a previously estimated disparity λ_s^{k-1} (coming from a coarser level $k-1$ or a previous iteration) and a refinement $d\lambda_s = \{d\lambda_s^k,\ \mathbf{s} \in S\}$ to be estimated. Considering the brightness constancy assumption for the total displacement yields the following equation:

$$I_1^k(\mathbf{s}) - I_2^k(\mathbf{s} + \mathbf{n}_s^k + [\lambda_s^{k-1} + d\lambda_s^k]\mathbf{v}_s^k) = 0, \tag{5}$$

to be solved with respect to $d\lambda_s^k$.

This equation involves pyramids of images $I_i^k = \{I_i^k(s),\ \mathbf{s} \in S\}$, $k = 0, ..., K$, $i = 1, 2$, and pyramids of tangent and normal vectors $\mathbf{n}^k = \{\mathbf{n}_s^k,\ \mathbf{s} \in S\}$, $\mathbf{v}^k = \{\mathbf{v}_s^k,\ \mathbf{s} \in S\}$ with k spanning from K (the coarsest resolution) to 0 (the finest resolution). The image pyramids I_i^k are derived from the original images I_i^0 by successive Gaussian smoothing and regular subsampling by a factor of two in each direction.

As for pyramid $\mathbf{n}^k$ and $\mathbf{v}^k$, we consider fundamental matrices $\{F^k\}$, $k = 0, ..., K$ deduced for each level from the initial matrix F and a change of coordinates. More precisely, we have:

$$F^k = M^{kT} F M^k,$$

where $M^k = \mathrm{diag}(2^k, 2^k, 1)$ is the matrix associated with the considered change of basis involved in the pyramidal representation. The matrix F^k allows to compute $\mathbf{n}_s^k$ and $\mathbf{v}_s^k$ at resolution k for each position s.

To insure that the previously estimated disparity at level $k-1$ follows the current epipolar geometry at a given level k, λ^k is deduced by projecting disparity $d\lambda^{k-1}$ (with a multiplying factor of 2) onto the epipolar lines at level k (see Fig. 2).

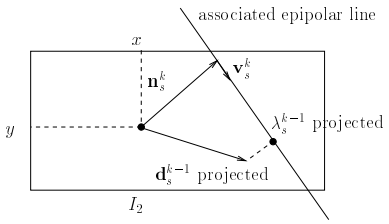

Fig. 2. *Projection of the disparity vector $d\lambda^{k-1}$ according to epipolar geometry at level k.*

2.3 Global estimation method

For sake of clarity, we will omit the resolution upper-script k in all expressions throughout the reminder of this paper. All the expressions will be meant to concern level k. Following the same principle as previously, the DFD expression (5) is linearized around point $\mathbf{s} + \mathbf{n}_s + \lambda_s \mathbf{v}_s$. This leads to a displaced version of the constrained optical flow equation:

$$d\lambda_s \mathbf{v}_s \nabla \tilde{I}_2(\mathbf{s}) + \tilde{I}_2(\mathbf{s}) - I_1(\mathbf{s}) = 0.$$

Assuming this equation is almost satisfied everywhere and that the disparity field is piecewise smooth, the disparity estimation problem may be addressed by the following minimization problem:

$$\widehat{d\lambda} = \arg \min_{d\lambda \in \mathbb{R}^{|\mathbf{s}|}} H(d\lambda) \ = \ \arg \min_{d\lambda \in \mathbb{R}^{|\mathbf{s}|}} (H_1(d\lambda, I_s) + \alpha[H_2(d\lambda)]), \qquad (6)$$

where α is an arbitrary fixed constant. The first term H_1 of the objective function H represents the data model:

$$H_1(d\lambda, I_s) = \sum_{\mathbf{s}\in S} \rho\left[d\lambda_s \mathbf{v}_s \nabla \tilde{I}_2(\mathbf{s}) + \tilde{I}_2(\mathbf{s}) - I_1(\mathbf{s})\right], \tag{7}$$

where $\tilde{I}_2 \stackrel{\triangle}{=} \{I_2(\mathbf{s} + \mathbf{n}_s + \lambda_s \mathbf{v}_s),\ \mathbf{s} \in S\}$ is the backward registered version of the second image. The second prior H_2 is the smoothness term which favors piecewise smooth disparity solutions. This term is expressed over all pairs $<s, r> \in \mathcal{C}$ of mutual neighbors (according to a 4-neighborhood system in our implementation):

$$H_2(d\lambda) = \sum_{<s,r>\in\mathcal{C}} \rho(||\mathbf{d}_s - \mathbf{d}_r||).$$

To cope with large deviations from the data model (resp. to allow disparity depth discontinuities), H_1 (resp. H_2) includes a M–estimator, ρ. Under some simple conditions [3, 5], (mainly the concavity of $\phi(v) \stackrel{\triangle}{=} \rho(\sqrt{u})$), any multidimensional minimization of the form "find $\arg\min_{x_i} \sum_i \rho(g_i(x))$" is equivalent to an optimization problem of the form "find $\arg\min_{x,z_i} \sum_i \tau z_i g_i(x)^2 + \psi(z_i)$" involving auxiliary variables (or weights) z_i's continuously lying in $(0, 1]$. The function ψ (which is never used in practice) is a decreasing function depending on ρ. In our case, the weights are of two natures: (a) the data outliers weights, $\delta = \{\delta_s,\ \mathbf{s} \in S\}$ (provided by the semi-quadratic formulation of H_1, and (b) the discontinuity weights $\beta = \{\beta_{sr},\ <s, r> \in \mathcal{C}\}$ related to the semi-quadratic formulation of H_2 and lying on the dual edge grid. The estimation problem is now expressed as a global minimization in $(d\lambda, \beta, \delta)$ of $\mathcal{H} = \mathcal{H}_1 + \alpha\mathcal{H}_2$ where:

$$\begin{cases} \mathcal{H}_1 = \sum_{s\in S} \tau_1 \delta_s [d\lambda_s \mathbf{v}_s . \nabla \tilde{I}_2(\mathbf{s}) + \tilde{I}_2(\mathbf{s}) - I_1(\mathbf{s})]^2 + \psi(\delta_s) \\ \mathcal{H}_2 = \sum_{<s,r>\in\mathcal{C}} \tau_2 \beta_{sr} \| d\lambda_s \mathbf{v}_s + \lambda_s \mathbf{v}_s + \mathbf{n}_s - \mathbf{d}_r \|^2 + \psi(\beta_{sr}) \end{cases},$$

where $\mathbf{d}_r = (\lambda_r + d\lambda_r)\mathbf{v}_r + \mathbf{n}_r$. The scalar τ_i a parameter depending on the M–estimator chosen.

The energy contribution of a point $\mathbf{s}$ to $\mathcal{H}_1$ is thus weighted by a factor $\delta_s \in (0, 1]$: the larger the contribution, the smaller the weight. Similarly, each pair of neighbors $<s, r> \in \mathcal{C}$ contributes to $\mathcal{H}_2$ with a weight $\beta_{sr} \in (0, 1]$ depending on their displacement vector difference $\|\mathbf{d}_s - \mathbf{d}_r\|$. The larger the difference, the smaller the weight.

The resulting semi-quadratic minimization problem is conducted alternatively with respect to the different variables (here the scalar field $d\lambda$ and the two weight fields δ and β). The minimization with respect to weights are given in the following closed from [3, 5]:

$$\arg\min_{z_i} \sum_i z_i g_i(x)^2 + \psi(z_i) = \frac{\rho'(g_i(x))}{2g_i(x)}.$$

Now considering weights as being frozen, the minimization with respect to $d\lambda_s$ is a classical weighted quadratic problem solved using an iterative method. Using a Gauss-Seidel scheme, the local update $d\lambda_s^{(n)}$ at iteration n of the iterative solver is given by:

$$d\lambda_s^{(n)} = \frac{\alpha(-\lambda_s\bar{\beta}_s + \mathbf{v}_s.\boldsymbol{\omega}_s^{n-1}) - \delta_s\mathbf{v}_s.\nabla\tilde{I}_2(\mathbf{s}).I_t(\mathbf{s})}{\delta_s(\mathbf{v}_s.\nabla\tilde{I}_2(\mathbf{s})^2 + \alpha\bar{\beta}_s}, \tag{8}$$

where $\boldsymbol{\omega}_s^{n-1}$ is the weighted average of neighboring disparity vectors at iteration $n-1$ and $\bar{\beta}_s$ is the sum the spatial discontinuity variables between s and its neighbors.

Let us note that in case of long range disparity, an initial disparity field is necessary to avoid solutions corresponding to undesirable local minima. In our case, we consider an initialization derived from the interpolation of the initial matched points of interest used for the computation of the fundamental matrix. We have used here a bilinear interpolation based on a Delaunay triangulation. The resulting field is projected on the top level of the pyramid to provide an initial disparity field for the coarsest resolution level with respect to the associated epipolar geometry (projection on the associated epipolar lines (see fig. 2)).

3 Segmentation

As our final goal is to provide a 3D reconstruction of the scene easy to handle, we now introduce a segmentation method of the dense disparity field obtained at the previous step. The method we propose is based on a adaptive triangular mesh structure. The idea of our technique consists in recursively splitting an initial mesh until each triangular element corresponds to a 3D planar element. The associated splitting criterion is based on the homographic parametric model-description of the disparity field. It can be easily shown that, according to a pinhole camera model, the disparity associated with a planar surface projected respectively as Π_1 in the first image and Π_2 in the second image satisfies an homographic model. This model is linear using homogeneous coordinates. For sake of clarity, all the following expressions are meant to be expressed in homogeneous coordinates. The homographic model links two corresponding points $\mathbf{s}$ and $\mathbf{s}+\mathbf{d}_s$ of Π_1 and Π_2 with a 3×3 homogeneous homography matrix named H up to a scalar factor μ:

$$\forall \mathbf{s} \in \Pi_1, H\mathbf{s} = \mu(\mathbf{s}+\mathbf{d}_s).$$

The segmentation step we propose consists thus in triangulating the disparity map until the disparity vectors associated with each patch correspond to a single representative homographic model. An initial Delaunay triangulation is first performed by taking four arbitrary points near to the corner of the image. This triangulation is then refined until each triangle verifies a distance criterion between the dense estimation disparity and an homographic model estimated within the considered triangle.

3.1 Homography estimation

The homography estimation is performed using a method proposed by Robert and Faugeras [13]. The method relies on the epipolar geometry to efficiently estimate the homography matrix from three or more corresponding pairs of points.

For H to be consistent with epipolar geometry, the homogeneous symmetric matrix $F^T H + H^T F$ must be null. This leads to 6 homogeneous equations with unknowns h_{ij} (the coefficient of H). In our case, each point of a considered triangle T accounts for one scalar equation. we have therefore the following system of equation:

$$\forall \mathbf{s} \in T,\ [\mathbf{s} + \mathbf{d}_s, F\mathbf{s}, H\mathbf{s}] = 0,$$

where $[a, b, c]$ denotes the triple product.

This over-constrained system can be rewritten in matrix notation as $A\mathbf{h} = 0$, where $\mathbf{h}$ is a 8 components vector gathering the unknown coefficients of H and A is a $(|\{\mathbf{s} \in T\}| + 6) \times 8$ matrix. An estimate of $\mathbf{h}$ is computed using a SVD (singular value decomposition) of the matrix $A^t A$.

To be robust to problematic situations where the estimated disparities are likely to be biased or erroneous (such as occlusion areas or range discontinuities), we exclude from this system points which are not simultaneously in accordance with the data model and the smoothing model (points for which the data outliers and the discontinuity weights approach zero).

3.2 Splitting criterion

The distance criterion we chose to handle the splitting of the triangular mesh is decomposed in two terms:

- The first one measures the adequacy of H to the disparity field. The influence of each point $\mathbf{s}$ of the triangle is weighted by the data model weight $\delta_{\mathbf{s}}$ coming from the robust estimator associated with the data model of the dense disparity estimator (occlusion areas do not influence the distance measurement). The resulting adequacy term is given by:

$$C_1(T, H, \mathbf{d}) = \frac{1}{\sum_{\mathbf{s} \in T} \delta_{\mathbf{s}}} \sum_{\mathbf{s} \in T} \delta_{\mathbf{s}} [\|H\mathbf{s} - (\mathbf{s} + \mathbf{d}_s)\|^2 + \|H^{-1}(\mathbf{s} + \mathbf{d}_s) - \mathbf{s}\|^2], \tag{9}$$

  where $\|\ \ \|$ denotes the Euclidean distance.
- The second term is related to the presence of disparity discontinuities within the considered triangle. This term is defined as the mean of discontinuity weights included in the considered triangle. It is expressed as follows:

$$\begin{cases} C_2(T, \beta) = \frac{\sum_{<\mathbf{s},\mathbf{r}> \in \mathcal{C}_T} \beta_{\mathbf{s},\mathbf{r}}}{|\mathcal{C}_T|} \\ C_T \triangleq \{<\mathbf{s}, \mathbf{r}>, \mathbf{s} \in T, \mathbf{r} \in T\}, \mathcal{C}_T \subset \mathcal{C} \end{cases} \tag{10}$$

  where $<\mathbf{s}, \mathbf{r}>$ denotes neighboring pixel of image 1.

More precisely, a given triangle T is split if the global criterion $C_1(T, H, \mathbf{d}) + \gamma C_2(T, \beta)$ exceeds a given threshold ϵ. The parameter γ is an arbitrary fixed positive constant.

3.3 Triangulation refinement

A given triangle is refined by adding a point located at its "center of mass"; the mass of each point being given by the value of their associated data outliers weights. The iterative triangulation refinement is performed until the distance measure $C(T)$ computed for each triangle decreases below ϵ.

4 3D reconstruction

So far we have obtained a 2D triangulation of the first image and an associated disparity information. The last step of our method consists in recovering 3D information in order to build the final 3D model. To that end, the calibration parameters of the cameras have to be estimated. As the aim is not an accurate reconstruction but a visually satisfactory 3D representation we have used a simplified self-calibration technique. This approach consists to fix to some arbitrary values the intrinsic parameters (represented by the A matrix) and then to estimate the extrinsic parameters. The intrinsic parameters are chosen in order to respect the following assumptions: the projection of the optical center is supposed to be at the center of the image, coordinate image axes are perpendicular, horizontal and vertical pixel sizes are fixed and equal to one and the focal length is assigned to a realistic value. The fundamental matrix F allows to access to the essential matrix E. This matrix only depends on the extrinsic parameters composed of the rotation matrix R and the translation vector between the first and the second camera location:

$$E = A^T F A = [\mathbf{t}]_\times R, \tag{11}$$

where $[\mathbf{t}]_\times$ is an antisymmetric cross product matrix associated to the translation vector $\mathbf{t}$. As shown by Tsai and Huang in [16], the essential matrix can be decomposed in order to recover rotation and translation parameters. Using a singular value decomposition, E can be written as follows:

$$E = \Delta E^{'} \Theta^t, \tag{12}$$

where Δ and Θ are two orthogonal matrices and $E^{'}$ is a diagonal matrix. It can be shown that an essential matrix has one null singular value, while the two others have the same value (they can be assigned to 1 because of the homogeneous property) [16]. Matrix $E^{'}$ can thus be rewritten as follows:

$$\begin{aligned} E^{'} &= T_1 R_1 \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}. \end{aligned} \tag{13}$$

Injecting this decomposition in equation (12) leads to write the matrix E as a product of an antisymmetric matrix by an orthogonal one. By identification with equation (11), we can extract rotation and translation matrices:

$$E = \underbrace{\Delta T_1 \Delta}_{[\mathbf{t}]_\times} \underbrace{\Delta^t R_1 \Theta^t}_{R}$$

We must notice that this decomposition is not unique. The rotation is only defined up to π while the translation is defined up to a scalar factor. The adequate pair of matrices is obtained by ensuring a 3D reconstructed point to be in view of the camera.

The resulting intrinsic and extrinsic parameters lead to two projection matrices. The nodes of the triangulation are then re-projected into the 3D space by solving the system (1) according to the dense disparity field. The 3D mesh is coded in the VRML language to allow real time interactivity. To avoid texture projection artifacts due to affine mapping, a preliminary simple correction is processed (an homographic transformation is performed to set the texture collinear to the image plane).

5 Results

The proposed method has been applied on different kinds of image sequences. It has been run both on real world sequences and synthetic sequences for which a ground truth exits.

The first sequence we are considering here is the well known synthetic "Yosemite sequence" (fig. 3). In order to satisfy the rigidity assumption, a major part of the sky containing moving clouds has been removed. Two different image pairs of this sequence have been considered. In the first one, which is composed of two consecutive images (images 9 and 10) the small range of the displacements (not more than 4 pixels) makes critical the estimation of the epipolar geometry. The second image pair, composed of far apart images in the sequence, constitutes a difficult benchmark towards the differential aspect of our method (up to 30 pixels of displacement).

As expected and shown on the recovered disparity map (fig. 3d), the disparities are larger in the mountain area in the foreground and continuously decreases while we move towards the valley. The global aspect of this map is in accordance with what could be expected from visual inspection.

Following [1], we provide quantitative comparative results on this pair of images. Angular deviations with respect to the actual flow field have been computed. Table 5 lists the mean angular value error and associated the standard deviation. It gathers some results presented in [1], and by other authors (only the higher and the lower mean square error obtained by state of the art methods are presented in comparison with the classical Horn and Schunck algorithm). Let us note, we report here only performances of similar algorithms (energy based dense estimators). Other results of more complex method combining motion estimation with a joint segmentation may be found in the literature. As may be

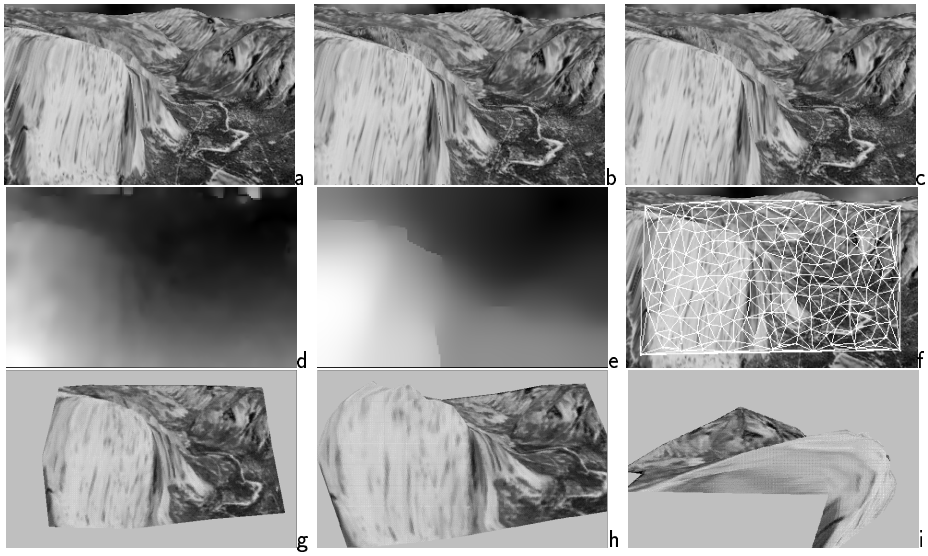

Fig. 3. Original images 3 (a), 11 (b) and 12 (c) of the "Yosemite" sequence; disparity map for images 11 and 12 (d) and 3 and 12 (e) (the darker the smaller the disparity value) and final triangulation (f); reconstructed images for translation along Z axis (small (g), important (h)) and for a complex motion (viewpoint on the left of the foreground mountain (i)

Technique	Mean error	Standard deviation
Horn and Schunck [1]	9.78^o	16.19^o
Black [2]	3.52^o	3.25^o
Lai and Vemuri[10]	1.99^o	1.45^o
Our method	4.82^o	3.27^o

Table 1. Comparative results on Yosemite

observed, compared to others our method yields to a higher angular discrepancy. Let us note meanwhile, it stays satisfactory. A few remarks must be done at this point. First at the opposite of the best methods mentioned in the table, our method uses a simple iterative solver (Gauss-Seidel). It could be therefore improved by using more efficient solvers. Second, it must be pointed out, that our method is a *one-dimentional method*. It is therefore far more faster than the others. Besides, due to small motion the epipolar geometry is quite difficult to estimate accurately.

Let us now consider the second sequence, composed of far apart images (images 3 and 12) of the "Yosemite" sequence. Experiments on this sequence have shown that, due to the presence of very large displacements (up to 30 pixels of displacement), non constrained optical flow estimators (even embedded in a multiresolution framework) do not converge towards acceptable solutions. As

shown in the disparity map presented figure 3e our method provides consistent results. The foreground mountain is characterized by important disparity values whereas in the background, disparities decrease smoothly. The dense disparity field estimation performs well for an image presenting both small and large displacements. The resulting field is globally smooth but presents discontinuities on important depth changes.

The disparity field computed from images 3 and 12 has been then iteratively triangulated to obtain a 2D model of the valley. The associated VRML model has been computed by arbitrarily fixing the focal length to 1000. Figure 3 presents some interpolated and extrapolated images. The camera displacement along the z-axis is not far away from the real 3D motion in images 3g and 3h. The resulting images are visually satisfactory. Image 3i exemplifies more complex displacements illustrating occlusion problems.

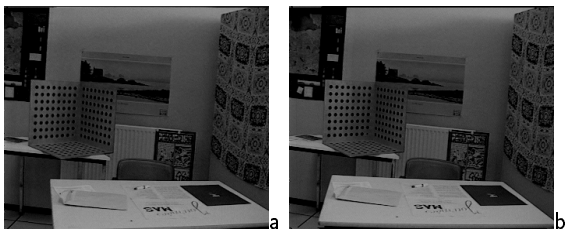

Fig. 4. Two original images of an indoor sequence (a) and (b)

Fig. 5. Three synthesized views from the same view point: model computed directly from automatically extracted and matched points (a, (b and (c; same 3D motion simulation as previously with the model obtained by our method ((d corresponds to (a, (e to (b and (f to (c)

Some reconstruction results obtained for a static scene shot by a moving commercial camera (fig. 4) are shown in figure 5. Two kind of reconstruction are presented here. The first one comes from the "image-matching" software, developed by Zhang [17], which gives a list of matching points of interest that respect the epipolar geometry. These points are triangulated and re-projected to obtain a 3D model. The examples presented in figure 5a 5b and 5c are constructed from 89 automatically extracted matching points. The synthesized views outline the presence of outliers points that make the model visually uncomfortable. This effect can mostly be explained by the presence of spurious matches that respect both the epipolar geometry and the luminance consistency. The second 3D model results from our algorithm (fig. 5d, 5e and 5f). A visual inspection of the reconstructed images shows far less artifacts for the same 3D displacements of the virtual camera. Such results could now be used in the context of video manipulation applications.

6 Conclusion

In this paper we have presented a method for the reconstruction of complex scene from a pair of uncalibrated images. This method relies on the estimation of a dense disparity field. The estimator proposed here is constrained by the epipolar geometry and incorporates robust function. We have experimentally demonstrated that the recovered fields are of good quality even in unfavorable case (very close views). The final 3D reconstruction is obtained through a segmentation process handled as a recursive adaptation of a triangular mesh. The outliers informations provided by the dense robust estimation are also used in the segmentation step to improve the quality of the final reconstruction. The efficiency of our approach has been validated on both polyhedral and non polyhedral complex scenes. The models obtained are sufficiently good to be used in a comfortable way in the context of video manipulation applications. Nevertheless, more accurate results could be expected using the best self-calibration method available in the literature. A natural extension of our algorithm would consist in considering the trifocal tensor (associated with three images [15]) instead of the fundamental matrix F, to avoid many degenerate estimation cases of F. This could naturally lead to take into account more than two images to improve the VRML model quality.

References

1. J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal on Computer Vision*, 12(1):43–77, 1994.
2. M. Black. Recursive non-linear estimation of discontinuous flow fields. In *Proceedings of the European Conf. on Computer Vision*, pages 138–145, Stockholm, Sweden, 1994.
3. M.J. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal on Computer Vision*, 19(1):57–91, July 1996.

4. B. Boufama and R. Mohr. Epipole and fundamental matrix estimation using the virtual parallax property. In *Proc. Int. Conf. Computer Vision*, pages 1030–1036, Cambridge, Massachusetts, 1995.
5. P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. on Image Processing*, 6(2):298–311, 1997.
6. O. Faugeras and S. Laveau. Representing three-dimensional data as a collection of images and fundamental matrices for image synthesis. In *Proceedings of International Conf. on Pattern Recognition*, pages 689–691, Jerusalem, Israel, 1994.
7. P. Havaldar, M-S. Lee, and G. Medioni. View synthesis from unregistered 2-d images. In *Graphics Interface*, pages 61–69, Toronto, Canada, May 1996.
8. T. Kanade, P.J. Narayanan, and P.W. Rander. Virtualized reality: Concept and early results. In *Workshop on Representation of Visual Scenes*, Cambridge, USA, 1995.
9. R. Koch, M. Pollefeys, and L. Van Gool. Automatic 3d model acquisition from uncalibrated image sequences. In *Computer Graphics International*, pages 597–604, Hannover, 1998.
10. S. Lai and B. Vemuri. Reliable and efficient computation of optical flow. *IJCV*, 29(2):87–105, 1998.
11. H.C. Longuet-Higgins. The reconstruction of a scene from two projections: configurations that defeat the 8-point algorithm. In *Proceedings of the 1st Conference on Artificial intelligence applications*, pages 395–397, Denver, 1984.
12. L. Robert and R. Deriche. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In *Proceedings of the 4th European Conf. on Computer Vision*, volume 1, pages 439–451, 1996.
13. L. Robert and O. Faugeras. Relative 3-D positioning and 3-D convex hull computation from a weakly calibrated stereo pair. *Image and Vision Computing*, 13(3):189–197, 1995.
14. P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, 1987.
15. P.H.S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.
16. R. Tsai and T. Huang. Uniqueness and estimation of three dimensional motion parameters of rigid objects with curved surface. *Transaction on Pattern Analysis ans Machine intelligence*, 6:13–26, 1984.
17. Z. Zhang. Estimating motion and structure from correspondences of line segments between two perspective images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1129–1139, 1995.