Generalization of Desargues Theorem for Sparse 3-D Reconstruction

Vincent Fremont, Ryad Chellali, Jean-Guy Fontaine

International Journal of Humanoid Robotics

GENERALIZATION OF DESARGUES THEOREM FOR SPARSE 3-D RECONSTRUCTION

VINCENT FREMONT
Heudiasyc UMR CNRS 6599
Universite de Technologie de Compiegne
BP 20529, 60205 Compiegne Cedex, France
vincent.fremont@hds.utc.fr

RYAD CHELLALI
Istituto Italiano di Tecnologia I.I.T.
Via Morego, 30 16163 Genova, Italy
ryad.chellali@iit.it

JEAN-GUY FONTAINE
Istituto Italiano di Tecnologia I.I.T.
Via Morego, 30 16163 Genova, Italy
jean-guy.fontaine@iit.it

Visual perception for walking machines needs to handle more degrees of freedom than for wheeled robots. For humanoids, four or six legged robots, camera motion is a 6-D one instead of 3-D or planar motions. Classical 3-D reconstruction methods cannot be applied directly because explicit sensor motion is needed. In this paper, we propose an algorithm for 3-D reconstruction of an unstructured environment using a motion-free uncalibrated single camera. Computer vision techniques are employed to obtain an incremental geometrical reconstruction of the environment and therefore using vision as a sensor for robots control tasks like navigation, obstacle avoidance, manipulation, tracking, etc. and 3-D model acquisition. The main contribution is that the off-line 3-D reconstruction problem is considered as a points trajectories search through the video stream. The algorithm takes into account the temporal aspect of the sequence of images in order to have an analytical expression of the geometrical locus of the points trajectories through the sequence of images. The approach is a generalization of the Desargues Theorem applied to multiple views taken from nearby viewpoints. Experiments on both synthetic and real image sequences show the simplicity and the efficiency of the proposed method. The method presented is that paper provides an alternative technical solution, easy to use, flexible in the context of robotic applications and can significantly improve the 3-D estimation accuracy.

Keywords: Computer Vision, Multi-views 3-D Reconstruction, Desargues Theorem, Images Sequence Analysis.

1. Introduction

3-D reconstruction from images is one of the main issues in computer vision. The use of video sensors provides both texture information at fine horizontal and vertical

resolution, and 3-D environment structure estimation from camera motion, which in turn enables navigation and obstacle avoidance in humanoids applications.

Basically, two main approaches have been investigated to solve the problem of structure from motion : short range motion-based methods and long range motion-based ones. In the first category, images are considered at distant time instants and a large camera displacement is generally performed to obtain accurate results. The images can be taken two by two to obtain the so-called *Stereovision Approach*[1,2] where the fundamental matrix can be used to constrain the matching and to make a projective reconstruction from disparity maps, three by three for the *Trifocal Approach*[3], and finally four by four for the *Quadrifocal Approach*[4]. These matching tensor techniques are essentially equivalent to implicit 3-D reconstruction methods and have two main drawbacks : the calibration procedure, which determines the intrinsic parameters of the sensor and its pose and orientation, and the inter-frame correspondence or features matching stage. If the first one can be solved[5], the second one is an ill-posed problem because hypothesis are made about the 3-D structure of the scene to help the inter-frame matching and it can only be solved using heuristic non linear optimization[6].

In the second category of reconstruction algorithms, images are considered at video rate (about 20-25 frames per second). Known techniques are *Optical Flow*[7] where the inter-frame apparent motion is estimated and sometimes combined with a measure of the camera velocity, *Factorization-based Reconstruction*[8,9] for which all the tracked image features are stacked in a matrix and solved by singular value decomposition to recover both motion and structure matrices, *Filter-design Approach*[10,11,12,13] where both structure and motion are estimated if the 3-D relative velocity between the scene and the camera is assumed constant over time, and finally *Hybrid Approach*[14,15] based on the use of the camera velocity estimate from odometry informations in the 3-D reconstruction process while relying on the tracking of 2-D sparse image features which is often called Simultaneous Localization and Mapping (SLAM). When there is no available odometric information, some approaches[2,13,16] are mainly based on 3D visual odometry, and use dense feature maps to get the position of the camera mounted on humanoid robots H-7 and HRP-2. It is important to notice that for these methods, the matching stage is less critical since the distance between two successive viewpoints is very small assuming image acquisition at video rate. But the 3-D reconstruction stage is also a global optimization problem mainly carried out on the re-projection error known as *Bundle Adjustment*[17,18], and a drift can appear from the accumulation of error during acquisition time.

The proposed method in this paper, is based on the generalization of the *Desargues Theorem*[19] using reference planes in an cluttered indoor environment. The algorithm takes into account the temporal aspect of the video stream in order to have an analytical expression of the geometrical locus of the points trajectories in a common reference plane through the sequence. Concerning plane-based 3-D reconstruction, one can quote the work of [20,21,22] where a plane has to be known

to estimate the motion and to constrain the 3-D reconstruction. Similar researches have been carried out by [23] using the Desargues theorem. Although a non linear optimization is used in our method to solve the 3-D reconstruction, the knowledge of an analytical form of the trajectories and of some temporal invariants, allows to simplify the matching stage[24]. Therefore the reconstruction quality is improved and above all, local minima, which are a major problem in global optimization methods, can be avoided.

The remainder of this paper is organized as follows. First we recall the geometrical model of image formation and the Desargues theorem used in this paper. In second we present the generalization of the Desargues Theorem for multi-view-based sparse 3-D reconstruction. In third we give the simulations and experimental results. We finally present the conclusion upon our work and further researches.

2. Preliminaries

2.1. *The camera model*

For the processing of the input images, the real camera have to be represented by a mathematical model. For that purpose and using the high quality of modern cameras, the imaging model considered here is the perspective projection model or pinhole camera model. From a mathematical point of view, the perspective projection[5] is represented by the projection matrix $\mathbf{P}$ of size 3×4 which makes corresponding to 3-D points in homogeneous coordinates $\tilde{\boldsymbol{X}} = [X, Y, Z, 1]^T$ the 2-D image points $\tilde{\boldsymbol{x}} = [u, v, 1]^T$:

$$\lambda \tilde{\boldsymbol{x}} = \mathbf{P}\tilde{\boldsymbol{X}} = \mathbf{K}[\mathbf{R}\ \boldsymbol{T}]\tilde{\boldsymbol{X}} \tag{1}$$

The (3×4) projection matrix $\mathbf{P}$ encapsulates the extrinsic parameters (camera position as (3×1) vector $\boldsymbol{T}$ and orientation as (3×3) orthogonal matrix $\mathbf{R}$) and the intrinsic parameters in (3×3) matrix $\mathbf{K}$ by considering that only the focal length f is unknown. The principal point $\boldsymbol{x}_0 = [x_0, y_0]^T$ is assumed to have the same coordinates as the image centre. It is also assumed square pixels, zero skew and no distortions which is quite realistic for modern video cameras.

2.2. *The Desargues configuration*

The 3-D reconstruction algorithm presented in this paper is based on a geometrical construction using the Desargues theorem which is on the basis of the projective geometry. By considering the Fig. 1, it states that[19] :

Theorem 1. *Let $[A, B, C]$ and $[A_1, B_1, C_1]$ be two triangles in the (projective) plane. The lines (AA_1), (BB_1) and (CC_1) intersect in a single point if and only if the intersections of corresponding sides (AB, A_1B_1), (BC, B_1C_1), (CA, C_1A_1) lie on a single line Δ.*

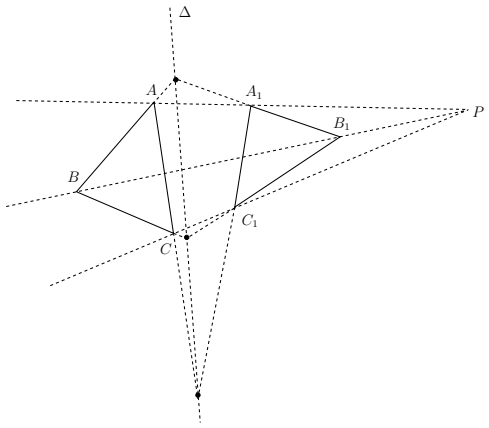


Fig. 1. Two triangles in Desargues configuration.

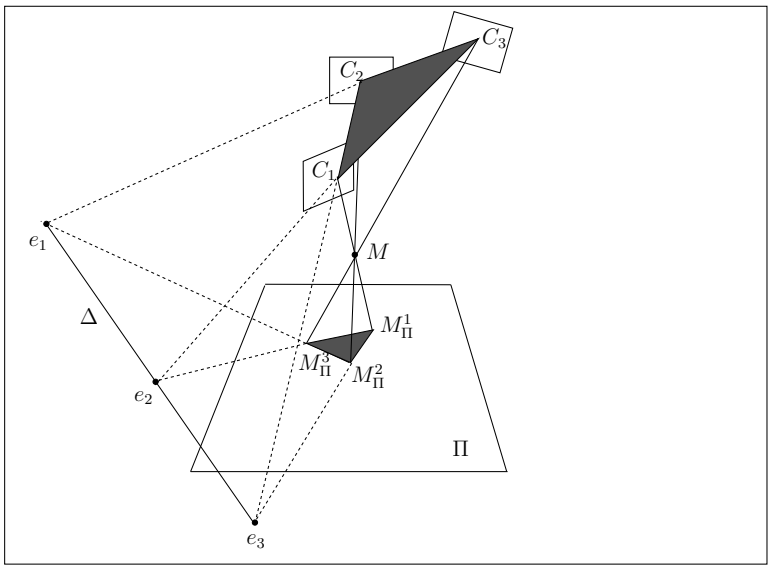


Fig. 2. Two triangles in Desargues configuration in the 3-D case.

The theorem has a clear self duality and express the *duality principle* of the projective geometry. This theorem, often used in 2D configuration, is also true in the 3-D case. Let the scene be composed with three camera represented by their image planes and their optical centers C_1, C_2 and C_3, and a reference plane Π (see Fig. 2).

Let M be a 3-D point in the observed scene. The projections of the camera optical centers C_i for $i = 1, \ldots, 3$ through M on the reference plane Π give respectively

the points M_Π^i for $i = 1, \ldots, 3$. The two triangles $[C_1, C_2, C_3]$ and $[M_\Pi^1, M_\Pi^2, M_\Pi^3]$ are in a Desargues configuration :

- The lines $(C_i M_\Pi^i)$ for $i = 1, \ldots, 3$ have for intersection the point M.
- The couples of lines $(C_i C_j)$ et $(M_\Pi^i M_\Pi^j)$ for $i, j = 1, \ldots, 3$ and $i \neq j$ intersect respectively in e_1, e_2 and e_3 which lie on the same line Δ on the reference plane Π.

Starting from that minimum configuration the next section describes the main step of the geometrical construction to find out the 3-D coordinates of the unknown point M for three views of the scene.

2.3. *3-D reconstruction for three views*

As we said previously, the two triangles $[C_1, C_2, C_3]$ and $[Tr_M]$ are in a Desargues configuration through M. Let P_i and P_j be two points which not belong to the reference plane Π (see Fig. 3). Let $[Tr_i]$ and $[Tr_j]$ be another triangles in the plane Π constructed like $[Tr_M]$ but through the reference points P_i and P_j.

By construction, the triangles $[C_1, C_2, C_3]$, $[Tr_i]$ and $[Tr_j]$ are in a Desargues configuration through P_i and P_j. It is easy to show that the triangles $[Tr_M]$, $[Tr_i]$ and $[Tr_j]$ are also in a Desargues configuration (see Fig. 2) because their corresponding sides intersections lie on the same line Δ (contained in Π). The intersection point of the lines joining the corresponding vertices of the triangles $[Tr_M]$ and $[Tr_i]$ is O_i and for $[Tr_M]$ and $[Tr_j]$ it is O_j.

But the points O_i and O_j are also the intersections between the 3-D lines (MP_i) and (MP_j) with the reference plane Π. Therefore the points O_i and O_j are invariant compare to the positions and orientations of the cameras referenced by their optical centers C_i for $i = 1, \ldots, 3$ since all triangles of the scene are in Desargues configuration.

Following that property, the two triangles $[Tr_i]$ and $[Tr_j]$ are also in a Desargues configuration and their vertices intersect in O, and that point is the intersection of the 3-D line (reference line) $(P_i P_j)$ with the reference plane Π. One more constraint is that by construction, O, O_i and O_j belong to the same line (see Fig. 3).

In conclusion, the 3-D coordinates of the unknown point M can be found out from the knowledge of the positions of the invariant points O_i and O_j in a reference plane Π and from two reference points (P_i and P_j).

3. Generalization : images sequence of N views

The previous case (three views configuration) can be applied to N multiple views. In this paper, images are considered at video rate therefore, the image flow can be considered as continuous in time. In the presence of a reference plane (here Π), the well-known linear plane-to-plane projective transformation can be used. The plane detection can be carried out by the algorithm proposed by [25].

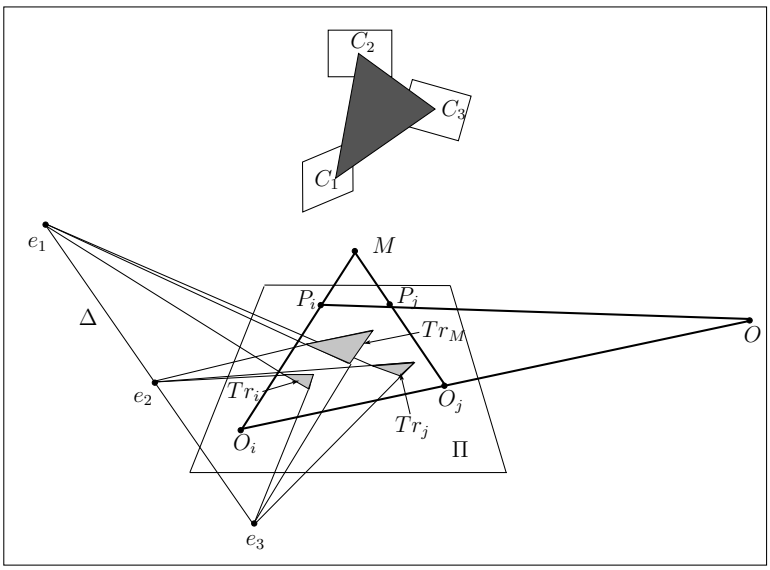


Fig. 3. Desargues configurations for the 3-D reconstruction of M.

Starting from the projection matrix of the perspective camera model (*e.g.* Eq. (1)), the 3-D points contained in the reference plane, for $Z = 0$ in the world reference coordinate system, it is possible to define the plane-to-plane (image plane to scene plane) transformation by its homography matrix $\mathbf{H}$. The homography matrix can be estimated using the proposed algorithm of [26].

The *homographic trajectory* $\tilde{\boldsymbol{x}}_h(t) = [u_h(t), v_h(t), 1]^T$ of the point to reconstruct $\tilde{\boldsymbol{M}} = [X, Y, Z, 1]^T$ is obtained by applying the inverse homography transformation to the image point trajectory :

$$\lambda_h(t) \begin{bmatrix} u_h(t) \\ v_h(t) \\ 1 \end{bmatrix} = \mathbf{H}^{-1}\mathbf{P}\tilde{\boldsymbol{M}} = \begin{bmatrix} 1 & 0 & \alpha_1(t) & 0 \\ 0 & 1 & \alpha_2(t) & 0 \\ 0 & 0 & \alpha_3(t) & 1 \end{bmatrix} \tilde{\boldsymbol{M}} \tag{2}$$

where the three parameters $\alpha_1(t)$, $\alpha_2(t)$ and $\alpha_3(t)$ are rational functions of the rotation matrix elements $\mathbf{R}$ and the translation vector $\boldsymbol{T}$. $\lambda_h(t)$ is an arbitrary scale factor due to the use of homogenous coordinates, and t is the time variable.

It is important to see that the focal length has been simplified, reducing the number of unknowns and permitting focus and de-focus during the image sequence acquisition. The homographic trajectories of image points corresponding to the 3-D points projected on the reference plane Π, are displayed in Fig. 4.

As we said previously, two 3-D reference points called $\boldsymbol{P}_i = [X_i, Y_i, Z_i]^T$ and $\boldsymbol{P}_j = [X_j, Y_j, Z_j]^T$ have to be known (eg. Fig. 3) to make the euclidean 3-D reconstruction possible. Their homographic trajectories are respectively $\boldsymbol{x}_i(t) = [u_i(t), v_i(t)]^T$ and $\boldsymbol{x}_j(t) = [u_j(t), v_j(t)]^T$. The point M to be reconstructed is on the

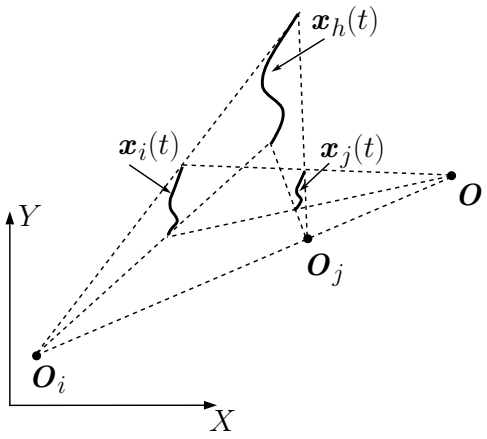


Fig. 4. Three points homographic configuration. It is important to notice that the three points O_x are invariants and aligned on the Desargues plane configuration.

3-D lines passing through the 3-D reference point P_i (or P_j) as it is shown in Fig. 3 and can be retrieved using the parametric equation of the 3-D lines (P_i, O_i) and (P_j, O_j) :

$$\begin{cases} X = X_i + \lambda_i(u_{0i} - X_i) = X_j + \lambda_j(u_{0j} - X_j) \\ Y = Y_i + \lambda_i(v_{0i} - Y_i) \quad = Y_j + \lambda_j(v_{0j} - Y_j) \\ Z = Z_i(1 - \lambda_i) \qquad\quad = Z_j(1 - \lambda_j) \end{cases} \tag{3}$$

The point with the coordinates $\boldsymbol{O}_i = [u_{0i}, v_{0i}]^T$ is the intersection point between the 3-D line (M, P_i) and the plane $Z = 0$ (called Π). In the same manner, $\boldsymbol{O}_j = [u_{0j}, v_{0j}]^T$ is the intersection point between the 3-D line (M, P_j) and the plane $Z = 0$. The Desargues configuration gives O_i and O_j to be invariant beside the camera motion. It can be seen that these two invariant points are only function of the 3-D coordinates of the two reference points (P_i and P_j) and the point M to be estimated :

$$\begin{cases} u_{0i} = \dfrac{X_i Z - X Z_i}{Z - Z_i} \ , \ v_{0i} = \dfrac{Y_i Z - Y Z_i}{Z - Z_i} \\ u_{0j} = \dfrac{X_j Z - X Z_j}{Z - Z_j} \ , \ v_{0j} = \dfrac{Y_j Z - Y Z_j}{Z - Z_j} \end{cases} \tag{4}$$

Therefore, in a first time, the following proposition can be stated :

Proposition 1. *If the homographic coordinates of* $\boldsymbol{O}_i = [u_{0i}, v_{0i}]^T$ *and* $\boldsymbol{O}_j = [u_{0j}, v_{0j}]^T$ *are known, it is possible to estimate the 3-D coordinates of* M *since it is the intersection of the 3-D lines* (P_i, O_i) *and* (P_j, O_j) *transformed in the homographic plane.*

A first estimation of the invariant points O_i and O_j over the N images of the sequence can be obtained by calculating the intersections of the lines (M_i, O_i) and (M_j, O_j) projected in the homographic plane (see Fig. 4) :

$$\begin{bmatrix} v_h(1) - v_i(1) & u_i(1) - u_h(1) \\ \vdots & \vdots \\ v_h(N) - v_i(N) & u_i(N) - u_h(N) \end{bmatrix} \begin{bmatrix} u_{0i} \\ \\ v_{0i} \end{bmatrix} = \begin{bmatrix} v_h(1)u_i(1) - u_h(1)v_i(1) \\ \vdots \\ v_h(N)u_i(N) - u_h(N)v_i(N) \end{bmatrix} \quad (5)$$

$$\begin{bmatrix} v_h(1) - v_j(1) & u_j(1) - u_h(1) \\ \vdots & \vdots \\ v_h(N) - v_j(N) & u_j(N) - u_h(N) \end{bmatrix} \begin{bmatrix} u_{0j} \\ \\ v_{0j} \end{bmatrix} = \begin{bmatrix} v_h(1)u_j(1) - u_h(1)v_j(1) \\ \vdots \\ v_h(N)u_j(N) - u_h(N)v_j(N) \end{bmatrix} \quad (6)$$

By using Eq. (2) and by injecting in it Eq. (3), the homographic trajectory of M becomes :

$$\begin{cases} u_h(t) = \dfrac{u_i(t)(1-\lambda_i) + \lambda_i \frac{u_{0i}}{\alpha_3(t)Z_i+1}}{1 - \lambda_i \frac{\alpha_3(t)Z_i}{\alpha_3(t)Z_i+1}} \\ \\ v_h(t) = \dfrac{v_i(t)(1-\lambda_i) + \lambda_i \frac{v_{0i}}{\alpha_3(t)Z_i+1}}{1 - \lambda_i \frac{\alpha_3(t)Z_i}{\alpha_3(t)Z_i+1}} \end{cases} \quad (7)$$

This couple of equations is still dependent to the temporal variable $\alpha_3(t)$. The two reference points P_i and P_j can be used to eliminate it. Setting back to the 3-D world, the parametric equation of the 3-D line (P_i, P_j) can be considered in the same manner as Eq. (3). The intersection point between the line (P_i, P_j) and the plane $Z = 0$ (see Fig. 3) is $\boldsymbol{O}$.

Since all parameters of the parametric equation of the 3-D line (P_i, P_j) are known, they can be injected in Eq. (2) and after some simple algebraic manipulations it is possible to express $\alpha_1(t)$, $\alpha_2(t)$ and $\alpha_3(t)$ as functions of the homographic trajectories $\boldsymbol{x}_i(t) = [u_i(t), v_i(t)]^T$, $\boldsymbol{x}_j(t) = [u_j(t), v_j(t)]^T$, the 3-D reference points P_i, P_j and the parameters of the 3-D line (P_i, P_j).

Finally a new time-dependant homographic trajectory of the point M to reconstruct can be obtained using the new formulations of $\alpha_1(t)$, $\alpha_2(t)$ and $\alpha_3(t)$ in Eq. (7). From that point, three other geometrical constraints have to be taken into account as depicted in Fig. 4 :

- the point $\boldsymbol{x}_h(t) = [u_h(t), v_h(t)]^T$, homographic trajectory of M, is the intersection of the 2D lines $\{(O_i, x_i(t)) \wedge (O_j, x_j(t))\}$, with $\wedge$ the intersection operator;
- the point O_i is the intersection between the 2D lines $\{(x_h(t), x_i(t)) \wedge (x_h(t + \Delta t), x_i(t + \Delta t))\}$ for two views taken at the instant time t and $t + \Delta t$;
- and by the Desargues configuration, the points O_i, O_j and O are aligned.

Using those intersection and alignment constraints into Eq. (7), one get the final analytical homographic trajectory $\boldsymbol{x}_h(t) = [u_h(t), v_h(t)]^T$ of M as rational functions of $\boldsymbol{x}_i(t)$, $\boldsymbol{x}_j(t)$ and $\boldsymbol{O}_j$ coordinates.

From the different homographic trajectories, two points are randomly selected and used to have a first estimation of the coordinates of the two points O_i and O_j using Eq. (5) and (6). Then the residuals are calculated using Eq. (8) in the following non linear criterion traducing the geometric distance from an estimated point $\hat{\boldsymbol{x}_h}(t)$ to a curve point $\boldsymbol{x}_h(t)$ over N images of the sequence :

$$\mathbf{J}(u_h(t), v_h(t)) = \sum_{t=0}^{N-1} (u_h(t) - \hat{u_h}(t))^2 + (v_h(t) - \hat{v_h}(t))^2 \tag{8}$$

with $[\hat{u_h}(t), \hat{v_h}(t)]^T$ the estimate of $[u_h(t), v_h(t)]^T$. From the LMeds estimator[27], the robust estimations of the points O_i et O_j can be obtained. The last step is to estimate the value of λ_i or λ_j, to obtain the 3-D euclidean coordinates of M. After some manipulations of Eq. (3), one obtains the following equation to get the value of λ_i :

$$\lambda_i = \frac{Z_j(u_{0j} - u_{0i})}{Z_i(u_{0j} - X_j) + Z_j(X_i - u_{0i})} \tag{9}$$

Therefore using λ_i, u_{0i} and v_{0i} in Eq. (3) we can get the 3-D coordinates of the unknown point M.

3.1. *The 3-D reconstruction algorithm*

The 3-D reconstruction algorithm developed in this paper can be summarized as follows :

(1) For each image of the sequence,
- Extract point features like corners or SIFT points using the algorithms presented in [28,29].
- Match the feature points from image motion[30] between two frames of the sequence of images.
- Calculate the homography matrix $\mathbf{H}$ using characteristic points in the reference plane.
- Transform each tracked feature points using the homography through the sequence to get the homographic trajectories.

(2) For each tracked feature point,
- Make m random draws on the homographic trajectories of $x_i(t)$, $x_j(t)$ and $x_h(t)$ to get a first estimate of $\hat{O}_i$ and $\hat{O}_j$ by solving Eq. (5) and (6) for each draw using for example a Singular Value Decomposition[31].

- Calculate the residual (Eq. (8)) and use the LMeds estimation algorithm to get $O_{i_{min}}$ and $O_{j_{min}}$.
- Calculate λ_i (or λ_j) using Eq. (9).
- Estimate the 3-D coordinates of M using Eq. (3).

3.2. *Singularities on camera motion and scene configuration*

Some singularities can appear in the case of particular camera motions and scene structures :

- If the camera motion is too weak compare to the dimension of the scene, there will not have enough random draws for robust estimation stage and the invariant points O_i and O_j will not be accurately estimated.
- If one of the reference point is on the same height ($Z_i = Z$) as the the point to estimate, the corresponding invariant point (O_i or O_j) will be rejected to infinity and will give numerical instabilities. The solution is to select a reconstructed 3-D point as a new reference point to estimate the current unknown 3-D point.

3.3. *Inter-frame features matching*

The method proposed by[24] uses Desargues configuration and properties to help the matching process. Indeed, as we stated before, the point $x_h(t)$, homographic trajectory of M, is the intersection of the 2D lines $(O_i, x_i(t))$ and $(O_j, x_j(t))$ where $O_i = (P_i, M) \wedge \Pi$ and $O_j = (P_j, M) \wedge \Pi$ in the 3-D scene.
Therefore, the matching process can be considered as the search of the points pairs O_i, O_j through the sequence which satisfy :

$$\begin{cases} (x_h(m), x_i(m)) \wedge (x_h(n), x_i(n)) = O_i \\ \\ (x_h(m), x_j(m)) \wedge (x_h(n), x_j(n)) = O_j \end{cases} \forall m, n \tag{10}$$

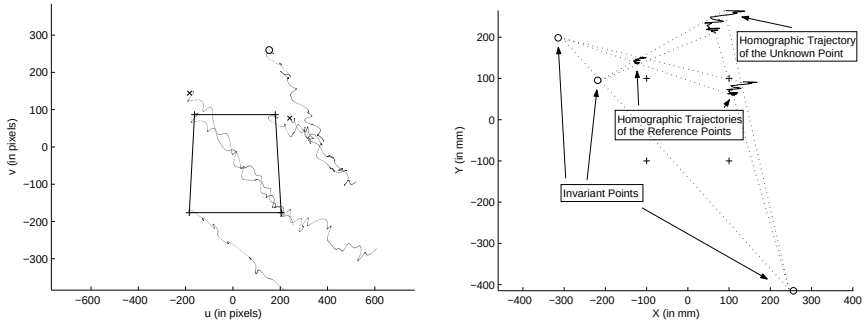
The matching algorithm extracts the $x_h(t)$ subsets that satisfy Eq. (10) by using a vote method where every point in the reference plane votes for a triplet (θ, H_i, H_j) using a multi-resolution approach. θ is the angle between the horizontal and the line containing the points O, O_i and O_j (see Fig. 4), and H_i, H_j are the euclidean distances from O_i, O_j to the horizontal line passing through O [24]. The triplet having the highest votes can be considered as subsets of corresponding points.

4. Simulations and Experimental Results

We present in this section the simulations and experimental results for the proposed 3-D reconstruction algorithm using the Desargues Theorem Generalization. Moreover the following hypothesis have to be considered : six known 3-D points are always visible during the sequence, four coplanar points to calculate the plane-to-plane homography and two known 3-D points, not contained in the reference plane, used as reference points.

4.1. *Experiments on synthetic image sequences*

A 3-D scene has been created with six known 3-D points and 100 unknown 3-D points randomly distributed in a cube of size $400 \times 400 \times 200$. The experiments on synthetic images have been carried out using Matlab[32] version 7.0.1. An example of the camera motion is shown in Fig. 4.1.a. using the image point trajectories. The camera is moving freely keeping the scene in its field of view. The 2-D homographic trajectories corresponding to three selected points are presented in Fig. 4.1.b. .



a. Points trajectories b. Homographic trajectories

The algorithm has been run on the image sequence for several levels of noise and the euclidean distances between 3-D real points and estimated ones are shown in Fig. 5.

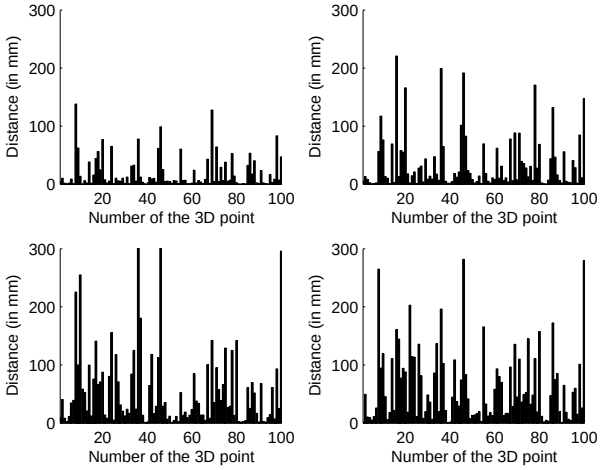
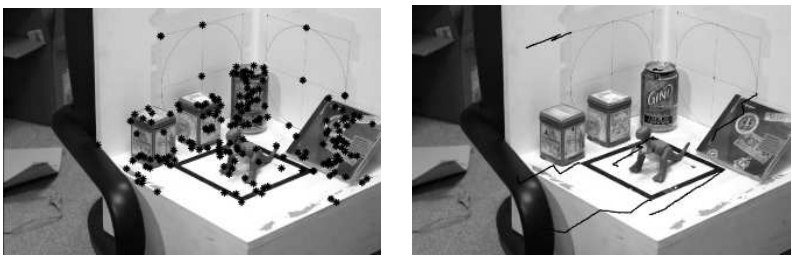


Fig. 5. Euclidean distance (in mm) between 3-D real points and estimated 3-D points for four levels of noise, from left to right and top to bottom : 2, 3, 4 and 5 pixels for the standard deviation

These simulations confirm the robustness of the proposed algorithm compared to noise level. The outliers are automatically rejected using using the LMeds estimation scheme. Nevertheless, bad estimations are still appearing, due to the configuration singularities as explained in Section 3.2.

4.2. *Experiments on real image sequences*

The first experimental system is composed of a digital camera Nikon D1. The Sequence 1 contains 147 images. The motion is complex enough with translations and rotations but stays planar. 174 feature points have been detected on the first image of the video (see Fig. 6.a.) using a combination of Harris corners and SIFT features.



a. Detected interest points on the first image

b. Tracked reference points

Fig. 6. Images of the 3-D scene presented in Sequence 1

Table 1. gives the 3-D estimations of the three points randomly selected in the scene.

Table 1. Real and estimated coordinates of the three points in the scene of Sequence 1.

	Real point (in mm)	Estimated point (in mm)
X_1	-28	-39.03
Y_1	171	169.58
Z_1	192	191.41
X_2	187	191.48
Y_2	-50	-48.27
Z_2	97	97.96
X_3	-62	-58.08
Y_3	102	117.76
Z_3	80	76.13

Let us point out that the reconstructed dimensions are correct but some coor-

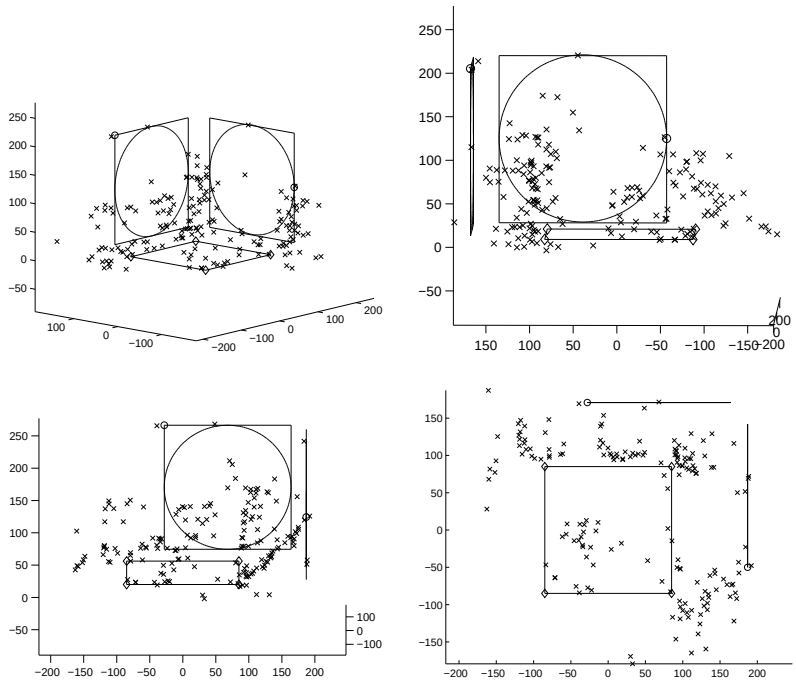


Fig. 7. Cloud of 3-D reconstructed points of the scene of Sequence 1. Legend : '◇' = known coplanar points, '○' = 3-D reference points, '×' = reconstructed points

dinates suffer from a lack of precision. The problem can be explained by the focus of the camera where some image regions become blur and the tracking algorithm is disrupted. Therefore the tracking window slides to a neighbor point while frame to frame SIFT descriptors are still good. Therefore the feature points trajectories become less precise through the sequence. On the other hand, if an estimated point is on the height as the reference point, a singularity can appear as stated in section 3.2.

Although a drift appears from the features tracking, geometrical constraints like coplanarity of the reconstructed shapes are still preserved.

Three other video sequences have be used to test the proposed algorithm. Two sequences have been filmed by a digital camera JVC KY-F1030U giving 7 images per seconds with a 1280 by 960 pixels of resolution. There are 166 images in the Sequence 2 and Sequence 3 is composed by 200 images. The scenes in the two last sequences are identical to Sequence 1 (coplanar points, reference points and same geometrical structures), but the objects in it and the camera motion are changing.

A large number of feature points (about 1000 for Sequence 2 and 500 for Sequence 3) have been detected on the first image of each video (see Fig. 8.a. and

Fig. 10.a.). Then the feature points have been tracked through image stream(about 330 points on the whole Sequence 2 and about 150 points on Sequence 3). The trajectories of the reference points have been superimposed to the first image of each sequence (see Fig. 8.b. and Fig. 10.b.).

a. Detected interest points on the first image

b. Tracked reference points

Fig. 8. Images of the scene used in Sequence 2

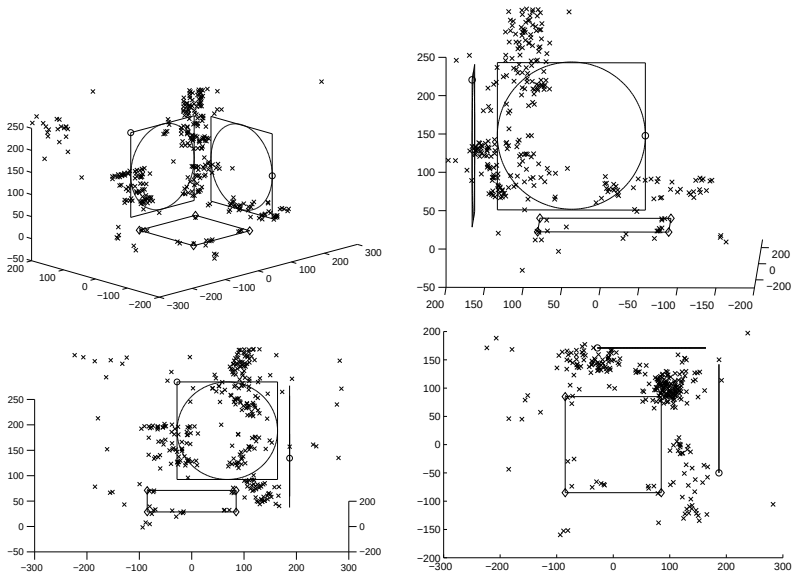


Fig. 9. Sets of reconstructed 3-D points for the scene of Sequence 2. Legend : '◇' = known coplanar points, 'o' = 3-D reference points, '×' = reconstructed points

Fig. 9 and 11 shows the cloud of 3-D points obtained with our 3-D reconstruction algorithm for four different viewpoints of the reconstructed scene. The different

a. Detected interest points on the first image

b. Tracked reference points

Fig. 10. Images of the scene used in Sequence 3

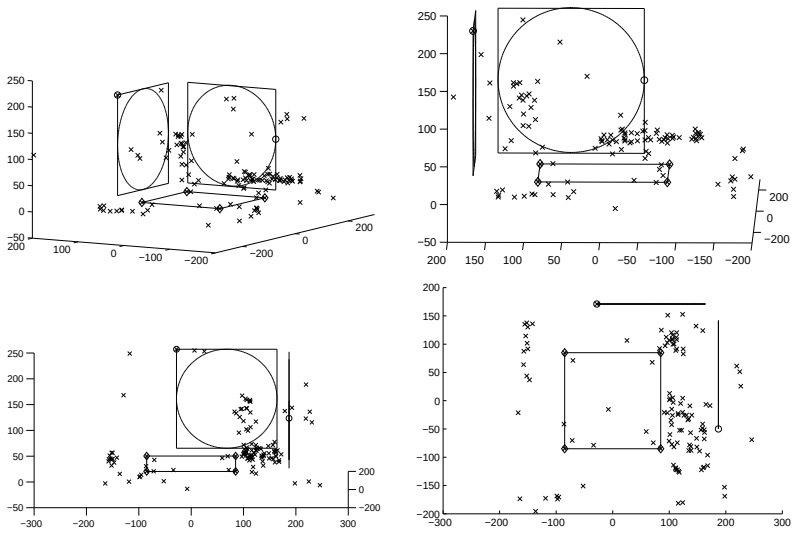


Fig. 11. Sets of reconstructed 3-D points for the scene of Sequence 3. Legend : '◇' = known coplanar points, '○' = 3-D reference points, '×' = reconstructed points.

objects of the scene and the dimensions have been well estimated. The coplanarity constraint is well preserved, as it is possible to see for the CD box or the electronic board. Taken as a whole, the 3-D positions of the sets associated to the various objects of the scene are well reconstructed and only the points in a singular configuration (same height or far from the reference points) are not well estimated.

In Sequence 3, the number of tracked points is less important, but the camera motion is more complex than the one in the second sequence. That last point insure

a good convergence of the algorithm and therefore a good 3-D reconstruction of the scene.

The last video, Sequence 4, has been filmed by a hand-waved digital camcorder JVC GR-DZ7E in PAL format ie. image resolution of 720 by 576 pixels at 25 frames per second. There are 1500 images (see Fig. 12 for samples) of a free walk in a corridor. Tracking of feature points has been made over all the images and then the video has been sub-sampled to keep only frames every 5 images for 3D reconstruction.

Fig. 12. Some key images from the video Sequence 4

The video Sequence 4 has been divided in three parts in order to propagate 3D points for an incremental 3D reconstruction. In each part, only the strongest feature points have been kept to have a sparse reliable reconstruction, which gives a total number of 730 points. The 3-D positions of the feature sets associated to the various objects of the scene are well reconstructed despite the low resolution of the video frames. The camera motion is totally free, which is suitable for humanoid robots purposes like works performed by Thompson et al.[2] and Ozawa et al.[16] on H-7 and HRP-2 robots. The reconstruction results as well as camera positions are shown in Fig. 13. One can notice the euclidean structure of the seen is preserved and the camera path is quite precisely reconstructed.

Table 2. gives the reconstruction times over the four sequences using a Pentium 4 3.2Ghz with 1 GB of RAM with Matlab version 7.0.1.

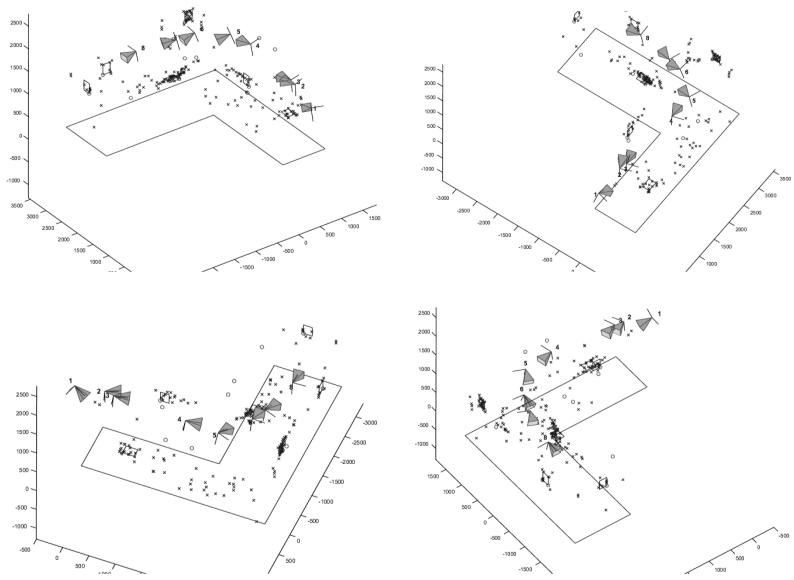


Fig. 13. Sets of reconstructed 3-D points for the scene of Sequence 4 with 8 key camera viewpoints. Legend : '$\circ$' = 3-D reference points, '$\times$' = reconstructed points, vertical rectangular shapes are used for homographies calculation and the large rectangular shape is the ground

Table 2. Reconstruction time for each images sequence.

	Sequence 1	Sequence 2	Sequence 3	Sequence 4
Nb. of points	174	1000	500	730
Nb. of images	147	166	200	300
Global time of reconstruction	5 min.	33 min.	9 min.	21 min.
Average time per point	6.5 sec.	7.7 sec.	3.42 sec.	5.37 sec.
Min. time per point	0.39 sec.	0.015 sec.	0.015 sec	0.015 sec.
Max. time per point	9.9 sec.	25.2 sec.	11.6 sec.	19.3 sec.

One can notice that the reconstruction time is increasing with the number of tracked points. This is mainly due to the robust estimation stage (not optimized), since the number of random subsets to be selected increases with the number of available feature points through the sequence. Moreover the use of Matlab routines increases greatly the computation time.

5. Conclusion and Further Research

We have described in this paper an original algorithm to get a sparse 3-D geometrical model of a scene from a video stream taken by a mobile monocular camera. The algorithm uses a generalization of the Desargues Theorem in the case of N multiple views taken at video rate. The temporal aspect of the video stream has been taken

into account in order to have an analytical expression of the geometrical locus of the image primitives through the video frames. The only constraint is to know some reference plane and two 3-D points for the first two images of the sequence. Then the estimated 3-D points can be used as new reference points and the environment can be reconstructed incrementally. Results on both simulated data and real videos are very encouraging and show the robustness and efficiency of the proposed algorithm.

Our approach can be used to construct an off-line 3D map for robots self-localization in an indoor environnement. It is important to notice that the computation time will be better using C/C++ code instead of Matlab routines.

Current work is devoted to the inter-frame matching stage as introduced in section 3.3. Since an analytical expression of the trajectories is available, further researches could be made in the sense of dense matching using time-filtering techniques to get a better visual aspect and more precise estimations of the reconstructed scene. Shape generation and texture-mapping techniques can also be used since the camera motion is known in order to get a more realistic 3-D model of the scene to introduce obstacle learning and avoidance.

Acknowledgements

The authors would like to thank the Institut de Recherche en Communications et en Cybernetique de Nantes, the Laboratoire de Vision et Robotique, France, for their support during the period 2002-2005, and all the people who were involved in this work.

References

1. M. Lhuillier and L. Quan, A quasi-dense approach to surface reconstruction from uncalibrated images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(3), 418–433 (2005).
2. S. Thompson and S. Kagami, Humanoid robot localisation using stereo vision, in *IEEE-RAS Int. Conf. on Humanoid Robots* (IEEE Press, Tsukuba, Japan, 2005), pp. 19-25.
3. Hartley, R. : Projective Reconstruction from Line Correspondence, in *IEEE Conference on Computer Vision and Pattern Recognition* (1994), pp. 903-907.
4. A.I. Comport, E. Malis and P. Rives, Accurate quadrifocal tracking for robust 3-D visual odometry, in *IEEE Int. Conf. Robotics and Automation (ICRA)* (IEEE Press, Roma, Italy, 2007), pp. 40-45.
5. O.D. Faugeras, Q.T. Luong and T. Papadopoulo (eds.), *The Geometry of Multiple Images*(MIT Press, Cambridge, Massachusetts, 2001), pp. 539–589.
6. Y. Ma, S. Soatto, J. Kosecka and S.S. Sastry, *An invitation to 3-D Vision*, 1st edn. (Springer-Verlag, New York, 2004), pp. 75–99.
7. K. Kanatani, N. Ohta and Y. Shimizu, 3-d reconstruction from uncalibrated-camera optical flow and its reliability evaluation, in *SCJ* **33**(9),1-10 (2002).
8. P. Sturm and B. Triggs, A factorization based algorithm for multi-image projective structure and motion, in *European Conference on Computer Vision* (2000), pp. 632-648.

9. C. Rother and S. Carlsson, Linear multi-view reconstruction and camera recovery using a reference plane, in *International Journal of Computer Vision* (2002).
10. L. Matthies, T. Kanade and R. Szeliski, Kalman Filter-based Algorithms for Estimating Depth from Image Sequences, in *International Journal of Computer Vision* (1989), pp. 209-236.
11. A. Chiuso, P. Favaro, H. Jin and S. Soatto, Structure from motion causally integrated over time, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(4)(2002), 523-535.
12. H. Jin, P. Favaro and S. Soatto, A semi-direct approach to structure from motion, in *The Visual Computer* (2003), pp. 1-18.
13. D. Chekhlov, M. Pupilli, W. Mayol-Cuevas and A. Calway, Robust Real-Time Visual SLAM Using Scale Prediction and Exemplar Based Feature Description. In *IEEE International Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, June 2007).
14. F. Chaumette, S. Boukir, P. Bouthemy and D. Juvin, Structure from controlled motion, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(5)(1996), 492-504.
15. O. Stasse, A.J. Davison, R. Sellaouti and K. Yokoi, Real-Time 3D SLAM for a Humanoid Robot considering Pattern Generator Information, in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*(2006).
16. R. Ozawa, Y. Takaoka, Y. Kida, K. Nishiwaki, J. Chestnutt, J. Kuffner, S. Kagami, H. Mizoguchi and H. Inoue, Using visual odometry to create 3D maps for online footstep planning, in *IEEE Int. Conf. on Systems, Man and Cybernetics (SMC2005)* (IEEE Press, Hawai, USA, 2005), pp. 2643-2648.
17. R. Mohr, L. Quan and F. Veillon, Relative 3-d reconstruction using multiple uncalibrated images, in *International Journal of Robotic Research* **14**(6)(1995), 619-632.
18. A. Heyden and K. Astrom, Euclidean reconstruction from image sequences with varying and unknown focal length and principal point, in *IEEE Conference on Computer Vision and Pattern Recognition* (1997), pp. 438-443.
19. R. Mohr and B. Triggs, Projective Geometry for Image Analysis, *INRIA Research Report* (1996).
20. Z. Zhang and A.R. Hanson, 3-D Reconstruction Based on Homography Mapping, in *ARPA96* (1996), pp. 1007-1012.
21. B. Triggs, Plane + parallax, tensors and factorization, in *European Conference on Computer Vision* (2000), pp. 522-538.
22. A. Bartoli and P. Sturm, Constrained structure and motion from n views of a piecewise planar scene, in *Conference on Virtual and Augmented Architecture* (2001), pp. 195-206.
23. A. Crimisi, I.D. Reid and A. Zisserman, Duality, rigidity and planar parallax, in *European Conference on Computer Vision* (1998), pp. 846-861.
24. R. Chellali, C. Maaoui and J.G. Fontaine, The Desargues Theorem to Build Matching Graph for N Images, in *2004 IEEE Conference on Robotics, Automation and Mechatronics* (2004).
25. M. Pollefeys, F. Verbiest and L.J. Van Gool, Surviving dominant planes in uncalibrated structure and motion recovery, in *European Conference on Computer Vision* (2002), pp. 837 ff.
26. K. Kanatani, N. Ohta and Y. Kanazawa, Optimal homography computation with a reliability measure, *IEICE Transactions on Information and Systems* **E83-D**(7)(2000), 1369-1374.
27. P. Rousseeuw and A. Leroy, in *Robust Regression and Outlier Detection* (John Wiley

and Sons, New York, 1987).
28. C. Harris and M.J. Stephens, A combined corner and edge detector, in *Alvey88* (1988), pp. 147-152.
29. D. G. Lowe, Distinctive image features from scale-invariant keypoints,*International Journal of Computer Vision* **60**(2), 91–110 (2004).
30. C. Tomasi and J. Shi, Good features to track, in *IEEE Conference on Computer Vision and Pattern Recognition* (1994).
31. P.E. Gill, W. Muray and M. Wright (edts), in *Practical Optimization* (Academic Press, 1989).
32. The MathWorks, http://www.mathworks.com/.

Biography

Vincent Fremont received his M.S. and Ph.D. degrees from the University of Nantes, in 2000 and 2003, respectively. From 2003 to 2005, he was Research and Teaching Associate at the ENSI de BOURGES. Now, he is a full time associate professor at the University of Technology of Compiegne and member of the CNRS Joint research unit Heuristic and Diagnoses Complex Systems.
His main field of interest deals with 3-D Computer vision, localization, pattern recognition and 3-D objects modeling.

Ryad Chellali received his degrees from Polytechnique Algiers and an Advanced Studies Diploma (DEA) in Robotics at Pierre et Marie Curie University (1988, Paris VI). He prepared his PhDs degree (1993) in Robotics at LRP (Robotics Lab of Paris) from the same University. In 2005, he received the Habilitation Diploma (D.Sc.) from the University of Nantes (Cybernetics Institute of Nantes). He served as Junior Researcher in 1992 at the French Institute of Transports (INRETS). From 1993 to 1995, he server as assistant professor at Pierre et Marie Curie University. From 1995 to 2006, he joined Ecole des Mines de Nantes, heading the automatic control chair. He is now senior researcher in the robotics, brain and cognitive sciences department of the Italian Institute of Technology.

His main field of interest deals with Telerobotics, virtual reality and human machine interfaces. Telepresence and Telexistence are also key words of his activity. Author of more than 50 papers in journals and conferences, he actively participated to over 10 European projects in the last 10 years. He received two French awards Creation of Innovative companies in 2000 and 2005 for his participation to start-up

companies launches.

Jean-Guy Fontaine received his degrees from ENSET and an Advanced Studies Diploma (DEA) in sensors and instrumentations at Orsay University (1983, Paris XI). He prepared his PhD's degree (1987) in Robotics at LIMRO from the same University. In 1995, he received the Habilitation Diploma (D.Sc.) from the prestigious Pierre et Marie Curie University (Paris VI) at L.R.P. (Laboratoire de Robotique de Paris). He served as Maitre de Conferences (Associate Professor) from 1987 to 1997 at Pierre et Marie Curie University heading an Electronic Chair. Then he accepted a full Professor Chair at ENSI de BOURGES, till 2006. He is now research director in the robotics, brain and cognitive sciences department of the Italian Institute of Technology.

His main field of interest deals with teleoperation, telepresence and telexistence with mobile robots (wheeled and legged machines). Micro and nano worlds exploration are also key words of his activity. Author of 7 patents, more than 120 papers in journals and conferences, he actively participated to over 35 European projects in the last 17 years with three success stories. Chevalier des Palmes Academiques of France, he had also honorific positions such as President of the French Robotic Association.