Real time registration of known or recovered multi-planar structures: application to AR

Gilles Simon, Marie-Odile Berger

Real time registration of known or recovered multi-planar structures: application to AR

Gille Simon and Marie-Odile Berger
LORIA - UHP Nancy 1 - INRIA Lorraine
{gsimon, berger}@loria.fr

Abstract

This paper describes an efficient and reliable method for real time camera tracking. Our algorithm makes use of a multi-planar model of the scene to achieve fast and accurate registration. In this paper, we also propose an automatic method for recovering the multi-planar structure of the scene directly from the homographies induced by the planes in the images. Results are presented, demonstrating tracking and reconstruction for indoor scenes. Videos of these results are available at http://www.loria.fr/˜gsimon/Bmvc.

1 Introduction

The objective of Augmented Reality (AR) is to add virtual objects to real video sequences, allowing computer-generated objects to be overlaid on the video in such a manner as to appear part of the viewed 3D scene. Applications of this concept concern interior design, architectural design, computer-aided repair and learning systems, medicine, and special effects for broadcast industry [1].

While there are several problems in building AR systems, one of the most basic challenge to overcome is the registration problem: the objects in the real and the virtual world must be properly aligned with respect to each other or the illusion that the two worlds coexist will be compromised. It is therefore essential to determine accurately the location of the cameras.

In this paper, we address the registration problem for interactive AR applications. Such applications require real-time registration process. Though the registration problem has received a lot of attention in the computer vision community, only little works have been devoted to sequential and real-time algorithms. In the present paper we propose a fast, reliable and sequential registration method designed for multi-planar environments. This method makes use of a model of planar structures which are visible in the scene. In this paper, we also show how to recover the planar structure of the scene from a set of snapshots.

2 Background

The current strategies for markerless AR can be divided into two main categories: model-based method and move-matching. Model-based techniques rely on the identification in the images of features from the object model. Pose estimation techniques can then be used to estimate the camera position [3, 8] on each single image. This capability of treating each image independently makes such methods more appropriate for real time implementations. Another consequence of model-based tracking is the absence of drift. However, such methods require significant manual intervention to construct the model.

Move-matching methods, also known as multi-frame structure-and-motion algorithms, appear to offer significant possibilities for general, accurate registration [5, 6]. Such systems simultaneously estimate camera motion and 3D structure of the imaged scene, by tracking key-points along the sequence. These systems permit accurate registration and negligible jitter. Besides the heavy calculation, these methods require a batch bundle adjustment in order to achieve a highly accurate registration [5]. Furthermore, the world-coordinate system is arbitrarily chosen, generally aligned with the first camera. As a result, the insertion of virtual objects requires further registration of object- to world-coordinate systems. These drawbacks preclude the sequential implementation required in autonomous applications.

Our approach combines both move-matching and model based methods. In this paper, we consider piecewise planar environments, like indoor scenes showing textured walls, or outdoor scenes including planar structures such as buildings and/or flat ground. If the model of the planar structure of the scene is available, we show how to recover the viewpoints from the homographies induced by the planes in the images. This work is an extension of [9] where the case of a single plane visible in the scene was addressed. It appears to be a good alternative for real-time AR because: 1. no marker is required ; 2. the matching of the key-points is constrained by the planar structure hypothesis and therefore is robust ; 3. no jitter sways out the augmenting virtual object ; 4. it performs in real-time, owing to the existence of a closed-form solution to the n-planes registration problem. Another contribution of this paper is to propose an automatic method for recovering the planar structure of the scene directly from the images.

The paper is organized as follows: section 3 deals with the 1-plane registration problem, section 4 extends it to n planes. Section 5 shows how the planar structures of the scene can be recovered when no model is available. Finally, experimental results are shown in section 6.

3 Single-plane registration

A single-plane temporal registration system was described in [9]. In this section, we remind how to compute a 3 × 4 projection matrix P^i in image i, from P^{i-1} and a planar homography H^i between these images.

We consider the pinhole camera model, which associates a point $\mathbf{x}^i$ in image i to a point $\mathbf{X}$ in the scene:

$$\mathbf{x}^i = \mathsf{P}^i\mathbf{X} = \mathsf{K}\left[\begin{array}{c} \mathsf{R}^i|\mathbf{t}^i \end{array}\right]\mathbf{X} \tag{1}$$

The matrix K represents the internal calibration parameters of the camera which are supposed to be known. $\left[\begin{array}{c} \mathsf{R}^i|\mathbf{t}^i \end{array}\right]$ is the viewpoint matrix to be estimated.

Let us now restrict $\mathbf{X}$ to lie on a plane Π and suppose that we know the associated planar homography H^i between image i and image $i-1$. The following relation is valid for all points on plane Π:

$$\mathbf{x}^i \propto \mathsf{H}^i\mathbf{x}^{i-1}, \tag{2}$$

where $\propto$ denotes an equality up to a multiplicative factor. Let $\mathtt{M}$ be a transformation matrix between $\mathcal{R}_0$, a coordinate system where the equation of Π is $z = 0$, and the world-coordinate system. For all points $\mathbf{X}$ on plane Π, we have:

$$\mathbf{x}^i = \mathsf{P}^i\mathbf{X} = \mathsf{P}^i\mathsf{M}\begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ 0 \\ 1 \end{pmatrix} = \left\langle \mathsf{P}^i\mathsf{M} \right\rangle \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ 1 \end{pmatrix} \tag{3}$$

where $\langle \mathsf{A} \rangle$ denotes the matrix A deprived of its third column. $\left\langle \mathsf{P}^i\mathsf{M} \right\rangle$ is invertible unless plane Π goes through the origin of the camera.

As a result, combining equations (2) and (3) yields:

$$\left\langle \mathsf{P}^i\mathsf{M} \right\rangle \propto \mathsf{H}^i \left\langle \mathsf{P}^{i-1}\mathsf{M} \right\rangle \tag{4}$$

Depriving $\mathsf{P}^i\mathsf{M}$ of its third column does not prevent from recovering the full viewpoint parameters. Indeed, knowing $\left\langle \mathsf{P}^i\mathsf{M} \right\rangle$ from equation (4), as well as the internal parameters K leads to:

$$\mathsf{K}^{-1} \left\langle \mathsf{P}^i\mathsf{M} \right\rangle \propto \left[\ \mathbf{r}_1\mathbf{r}_2\mathbf{t}\ \right] \tag{5}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are orthonormal vectors. The third column for the rotation matrix of the viewpoint is merely given by $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$. In practice, the orthonormality conditions are never perfectly met, and renormalization must be applied ($\mathbf{r}_2 = \mathbf{r}_2/||\mathbf{r}_2||, \mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2/||\mathbf{r}_1 \times \mathbf{r}_2||, \mathbf{r}_1 = \mathbf{r}_2 \times \mathbf{r}_3$).

4 Multi-plane registration

In this section, we suppose that $n > 1$ homographies H_p^i are known for n planes Π_p. From equation (4), we get for each plane:

$$\left\langle \mathsf{P}^i\mathsf{M}_p \right\rangle \propto \mathsf{H}_p^i \left\langle \mathsf{P}^{i-1}\mathsf{M}_p \right\rangle. \tag{6}$$

Three methods can be used to compute P^i from the set of equations (6). We first show how to estimate iteratively the viewpoint parameters. Then we present a linear solution for the computation of P^i. Finally, we propose a linear estimator of the viewpoint, using an approximation of the rotation matrix.

Iterative computation of the viewpoint (ITER)

For each correspondence $\mathbf{x}_j \leftrightarrow \mathbf{x'}_j$ between image $i-1$ and image i, we have

$$\mathbf{x'}_j \propto \mathsf{H}_p^i\mathbf{x}_j, \tag{7}$$

where p is the subscript of the plane containing the related 3D point $\mathbf{X}_j$. A residual r_j can hence be computed for each pair of points: $r_j = dist\left(\mathsf{H}_p^i\mathbf{x}_j, \mathbf{x'}_j\right)$, where $dist$ is the Euclidean distance between two 2D points expressed in homogeneous coordinates. From (6), we get

$$\mathsf{H}_p^i \propto \left\langle \mathsf{P}^i\mathsf{M}_p \right\rangle \left\langle \mathsf{P}^{i-1}\mathsf{M}_p \right\rangle^{-1}. \tag{8}$$

Consequently, the residuals expressively depend on the 6 parameters of the viewpoint in image i (translation $[t_x t_y t_z]^\top$ and Euler angles α, β, γ). These parameters can be recovered through a least squares minimization:

$$\min_{t_x, t_y, t_z, \alpha, \beta, \gamma} \sum_j r_j^2 \tag{9}$$

with P^{i-1} providing the initial parameters.

This method has proven to be stable and accurate in our experiments. However, the iterative process is relatively slow (typically 0.5 sec per frame in the conditions of the tests presented in section 6.1), and does not permit real time on common computers.

Linear computation of the projection (LIN1)

This section takes its inspiration from the reasoning used in [6] for the linear computation of a homography. From equations (7) and (8), and considering that $\left\langle \mathsf{P}^i \mathsf{M}_p \right\rangle = \mathsf{P}^i \left\langle \mathsf{M}_p \right\rangle$, we obtain for each correspondence $x_j \leftrightarrow x'_j$:

$$\mathbf{x'}_j \propto \mathsf{P}^i \left\langle \mathsf{M}_p \right\rangle \left\langle \mathsf{P}^{i-1} \mathsf{M}_p \right\rangle^{-1} \mathbf{x}_j \Leftrightarrow \mathbf{x'}_j \propto \begin{pmatrix} \mathbf{p}^{1\top} \mathsf{Q}_p \mathbf{x}_j \\ \mathbf{p}^{2\top} \mathsf{Q}_p \mathbf{x}_j \\ \mathbf{p}^{3\top} \mathsf{Q}_p \mathbf{x}_j \end{pmatrix}, \tag{10}$$

where $\mathbf{p}^{k\top}$ is the k^{th} row of matrix P^i, and $\mathsf{Q}_p = \left\langle \mathsf{M}_p \right\rangle \left\langle \mathsf{P}^{i-1} \mathsf{M}_p \right\rangle^{-1}$. Writing equations (10) in terms of cross products gives

$$\begin{pmatrix} y'_j \mathbf{p}^{3\top} \mathsf{Q}_p \mathbf{x}_j - w'_j \mathbf{p}^{2\top} \mathsf{Q}_p \mathbf{x}_j \\ w'_j \mathbf{p}^{1\top} \mathsf{Q}_p \mathbf{x}_j - x'_j \mathbf{p}^{3\top} \mathsf{Q}_p \mathbf{x}_j \\ x'_j \mathbf{p}^{2\top} \mathsf{Q}_p \mathbf{x}_j - y'_j \mathbf{p}^{1\top} \mathsf{Q}_p \mathbf{x}_j \end{pmatrix} = \mathbf{0},$$

where $\mathbf{x'}_j = \begin{bmatrix} x'_j & y'_j & w'_j \end{bmatrix}^\top$. Finally, as we have $\mathbf{p}^{k\top} \mathsf{Q}_p \mathbf{x}_j = \mathbf{x}_j^\top \mathsf{Q}_p^\top \mathbf{p}^k$, we get the linear system of equations:

$$\begin{bmatrix} \mathbf{0}^\top & -w'_j \mathbf{x}_j^\top \mathsf{Q}_p^\top & y'_j \mathbf{x}_j^\top \mathsf{Q}_p^\top \\ w'_j \mathbf{x}_j^\top \mathsf{Q}_p^\top & \mathbf{0}^\top & -x'_j \mathbf{x}_j^\top \mathsf{Q}_p^\top \\ -y'_j \mathbf{x}_j^\top \mathsf{Q}_p^\top & x'_j \mathbf{x}_j^\top \mathsf{Q}_p^\top & \mathbf{0}^\top \end{bmatrix} \begin{pmatrix} \mathbf{p}^1 \\ \mathbf{p}^2 \\ \mathbf{p}^3 \end{pmatrix} = \mathbf{0}. \tag{11}$$

Although there are three equations, only two of them are linearly independent (we omit the third equation). Therefore, each correspondence that belongs to one plane Π_p gives two equations in the entries of P^i. Finally, we obtain a linear system of equations having the form $\mathsf{A}\mathbf{p} = \mathbf{0}$, where A is a $2m \times 12$ matrix, m being the number of point correspondences. This system may be solved very quickly using a Singular Value Decomposition (SVD).

Linear approximation of the viewpoint (LIN2)

The above method linearly computes the 12 entries of P^i. However, no profit is taken from the a priori knowledge of the matrix K and that R is a rotation matrix. Therefore, this method is in practice less accurate than the iterative method (see section 6.1).

To overcome this problem, we suppose that the camera rotation between two images is small. Thus we can perform a first order approximation of this rotation $\Delta\mathsf{R}^i(\Delta\alpha, \Delta\beta, \Delta\gamma)$, where $\Delta\mathsf{R}^i = \mathsf{R}^i(\mathsf{R}^{i-1})^t$. We obtain a linear expression of the entries of P^i in the coefficients $t_x, t_y, t_z, \Delta\alpha, \Delta\beta, \Delta\gamma$:

$$\mathsf{P}^i \approx \mathsf{K} \left[\begin{bmatrix} 1 & -\Delta\alpha & \Delta\beta \\ \Delta\alpha & 1 & -\Delta\gamma \\ -\Delta\beta & \Delta\gamma & 1 \end{bmatrix} \mathsf{R}^{i-1} \quad \begin{matrix} t_x \\ t_y \\ t_z \end{matrix} \right]. \tag{12}$$

Combining equations (11) and (12) provides a linear system in the viewpoint parameters.

Overview of the system
Finaly, the main steps of this algorithm are summarized below:

Initialization stage:

1. Give or acquire the equation of the observed planes used for registration,
2. Compute the projective matrix for the first frame P^0,

Computation of the projective matrix P^i for $i > 0$:

1. Compute the set of matched key-points between images $i - 1$ and i for each observed plane.
2. Compute the homographies induced by the planes between $i - 1$ and i.
3. Compute P_i from P_{i-1} and the computed homographies

5 Recovery of the planar structures

The methods described above supposed that a multiplanar model of the scene is known. When this knowledge on the scene is not available, we show in this section how to recover the 3D parameterization of the observed planes in an off-line process by using planar structure-and-motion.

Ideally, an AR system should work in all environments and the user should walk anywhere he pleases. On the contrary , during the modeling stage, we can impose more restrictive hypothesis in order to make easier and to robustify the modeling stage: we can impose that the intrinsic camera parameters are constant and are computed before the acquisition process. We can also consider good quality sequences... Our aim is to obtain an automatic method for scene structure recovery which can be done in reasonable time just before the application.

Our algorithm for recovering the Euclidean structure of the scene using a planar structure-and-motion algorithm is inspired from recent works in the field [7, 10, 2]. These works have in common to use the homographies induced by several planes in two or more images to recover the structure of the scene as well as the camera motion: most of the time, a first estimate is obtained from the computed epipoles [10, 2]; then bundle adjustment techniques are used to refine the structure of the scene.

In what follows, we propose a simple and fast solution which takes advantage of the known camera parameters. First, an initial guess for the structure and motion algorithm can be obtained from any homography induced by one of the plane in the scene using the single-plane algorithm. These estimates are then refined by minimizing an homographic cost function over the sequence or the considered snapshots. Note that we do note minimize the classical residual $\sum d(Hx_i, x'_i)$ for at least two reasons: (i) as non linear minimization must be performed, such a distance requires heavy computational time due to the number of matched points. (ii) it is often useful to be able to perform structure and motion from the homographies themselves instead than from a set of matched point. For distant images, only a small set of points can be matched between such views. On the contrary, the homography induced by a plane can be easily computed from far views by compositing frame to frame homographies.

We first show how to recover structure and motion using two perspective images of n planes. Then we extend the method to the case of q views. Let P^1 and P^2 be the projection matrices computed for the two frames. The planar homography for plane Π_p between these two views is denoted by H_p. One of the observed plane is used to recover the viewpoint using the single-plane method. This plane is called the reference plane in the following and is denoted by Π_1. As the internal parameters are constant over time $\mathsf{P}^1 = \mathsf{K}[\mathsf{R}^1|\mathbf{t}^1]$ and $\mathsf{P}^2 = \mathsf{K}[\mathsf{R}^2|\mathbf{t}^2]$. By construction, the projection matrix is expressed in a frame attached to the reference plane. Expressing these matrices in the first camera frame, we obtain the canonical representation: $\mathsf{P}^1 = \mathsf{K}[\mathtt{I}|\mathtt{0}]$ and $\mathsf{P}^2 = \mathsf{K}[\mathsf{A},\mathbf{a}]$ where $\mathsf{A} = \mathsf{R}^2\mathsf{R}^{1\top}$ and $\mathbf{a} = \mathbf{t}^2 - \mathsf{R}^2\mathsf{R}^{1\top}\mathbf{t}^1$.

The plane parameterization can be retrieved through the explicit expression of the induced homography given in [6]. Given the projection matrices $\mathsf{P}^1 = \mathsf{K}[\mathtt{I}|\mathtt{0}]$ and $\mathsf{P}^2 = \mathsf{K}[\mathsf{A},\mathbf{a}]$, and given a plane defined by $\pi^\top \mathbf{X}_c = 0$ with $\pi = (\mathbf{v}^\top, 1)^\top$, the homography H induced by this plane is:

$$\mathsf{H} \propto \mathsf{K}(\mathsf{A} - \mathbf{a}\mathbf{v}^\top)\mathsf{K}^{-1}, \tag{13}$$

or equivalently $\mathsf{K}^{-1}\mathsf{H}\mathsf{K} \propto \mathsf{A} - \mathbf{a}\mathbf{v}^\top$.

As the homography H_p is known for each plane Π_p, and if we suppose that the estimates of A and a given by the single-plane method is correct, the parameterization $\mathbf{v}_p$ of these planes can be easily computed using (13) which is linear in $\mathbf{v}_p$. Let $\mathbf{h}$ be the 9-vector made up of the entries of the matrix $\mathsf{K}^{-1}\mathsf{H}_p\mathsf{K}$ and $\mathbf{b}$ be the 9-vector made up of the entries of the matrix $\mathsf{A} - \mathbf{a}\mathbf{v}_p^\top$. Since $\mathsf{K}^{-1}\mathsf{H}_p\mathsf{K}$ and $\mathsf{A} - \mathbf{a}\mathbf{v}_p^\top$ are proportional, each cross-product $\mathsf{C}(\mathsf{A},\mathbf{a},\mathbf{v}_p)_{kl} = \begin{vmatrix} \mathbf{h}_k & \mathbf{b}_k \\ \mathbf{h}_l & \mathbf{b}_l \end{vmatrix}$ is null ($1 \leq k,l \leq 9$). Hence, a linear system of equations in the 3 $\mathbf{v}_p$ coordinates can be built. It is solved using its SVD.

As the viewpoint estimate given by the single plane method is not always accurate (see section 6), a global adjustment technique is then used to refine both the viewpoint $[\mathsf{A},\mathbf{a}]$ and the plane parameterizations $\mathbf{v}_p$ ($1 \leq p \leq n$). We consider the set of equations

$$\left\{ \begin{array}{ccc} \mathsf{H}_1 & \propto & \mathsf{K}(\mathsf{A} - \mathbf{a}\mathbf{v}_1^\top)\mathsf{K}^{-1} \\ & \ldots & \\ \mathsf{H}_n & \propto & \mathsf{K}(\mathsf{A} - \mathbf{a}\mathbf{v}_n^\top)\mathsf{K}^{-1} \end{array} \right.$$

Here, $\mathbf{v}_1$ is the parameterization of the reference plane. As the equation of this plane is $\mathtt{Z} = 0$ in the initial frame, the parameterization in the first camera frame is $\mathbf{v}_1 = -\mathsf{R}^1 k / k^\top \mathsf{R}^{1\top}\mathbf{t}^1$ where $k^\top = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$.

The set of unknowns $\mathbf{v}_2, \ldots, \mathbf{v}_n, \mathsf{A}, \mathbf{a}$ is finally tuned by iteratively minimizing, using Powell's algorithm, the following cost function:

$$Min_{\alpha,\beta,\gamma,a,v_2,\ldots,v_p} \sum_{p=1}^{n} \sum_{k,l=1}^{k,l=9} \mathsf{C}(\mathsf{A},\mathbf{a},\mathbf{v}_p)^2_{kl}$$

where α, β, γ are the Euler angles of A. To cope with the classical ambiguity between a and v (λa and $\frac{1}{\lambda}v$ is also solution), one of the component of a is fixed to its initial value during minimization. If more than two views (let us say Q) are considered, we have to determine Q canonical transformations $P^q = [A^q|a^q]$. We then have to minimize

$$Min_{\alpha^1,\ldots,\alpha^Q,\beta^1,\ldots,\beta^Q,\gamma^1,\ldots,\gamma^Q,a^1,\ldots,a^q,v_2,\ldots,v_p} \sum_{q=1}^{Q} \sum_{p=1}^{n} \sum_{k,l=1}^{k,l=9} \mathsf{C}(\mathsf{A}^q,\mathbf{a}^q,\mathbf{v}_p)^2_{kl}$$

6 Experimental results

Results are first presented on a calibration target sequence to asses the accuracy of our algorithm. Then, results obtained on a poorly textured indoor sequence are provided.

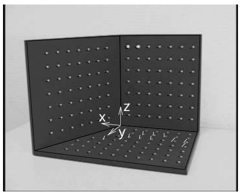
6.1 A target sequence

This sequence contains 98 images. The camera is moving around the calibration target, roughly pointing the intersection of the three planes (see Fig. 1). The initial matrix P^0 is obtained using classical calibration method in [4] from 3D/2D correspondences.

Comparison of the three methods ITER, LIN1, LIN2

Fig. 1(b) shows the temporal evolution of one viewpoint parameter (x translation), for each of the proposed methods. The other viewpoint parameters show similar curves and errors. Actual values (represented by crosses in Fig. 1), are computed every ten images, using classical calibration from points on the target. ITER gives rise to the best results, LIN1 to the worst. The graph obtained for LIN2 is relatively close to the actual graph, but it slowly diverges from it: this expresses an accumulation error due to the successive approximations of the rotation matrix. However, the pose estimation error obtained at the end of the sequence is only 6.3 cm for a distance target - camera equal to 127 cm. This error is almost not perceptible with regard to the projection of the target. The computation rates are detailed in table Fig. 1.b for each step of the algorithm LIN2 on a Pentium III 900 Mhz. The whole algorithm performs in 62.70 ms per frame on average, which leads to an approximate processing rate of 16 frames per seconds. The iterative method (ITER) is more time consuming.

a.



b.

MIC	30,33 ms
Matching	18,62 ms
RANSAC	12,13 ms
LIN2	1,62 ms
Total	62,70 ms

c.

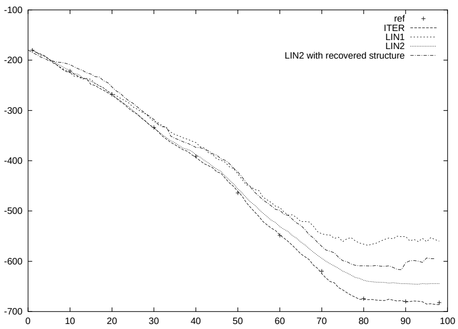


Figure 1: (a) One snapshot of the target sequence (b) Computation time (c) Temporal evolution of t_x with thee planes, for ITER, LIN1, LIN2 and LIN2 with the recovered structure. The crosses represent the actual values of t_x

Influence of the number of planes

The above results were obtained by registering three planes of the target. Figures 2 shows results obtained by using ITER (the most accurate method) with one or two planes. The titles of the curves indicate which planes were used: X represents the vertical left plane of the target (see Fig. 1), whose equation is $\mathrm{X} - 0.577\mathrm{Y} = 0$, Y the vertical right plane ($\mathrm{Y} = 0$) and Z the horizontal plane ($\mathrm{Z} = 0$). Fig. 2 shows that the results obtained from a single plane are much more irregular than the results obtained from two or three planes, which illustrates our contribution with regard to [9].

Recovery of the multi-planar structure

In this section we exhibit results of the recovering process. First we want to prove that the estimation given by the single plane method is good enough so that the global adjustment

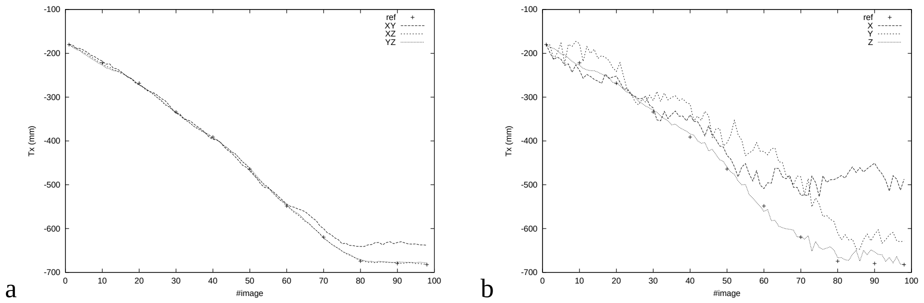


Figure 2: Temporal evolution of t_x when two (a) or one plane (b) are used.

process converges towards the actual solution. Fig. 3.a exhibits the recovered translation for each frame of the sequence. The initial estimates of t_x given by each of the visible planes (x-plane y-plane and z-plane) are shown in dotted lines. The results of the batch process is shown with bold lines (x_Tx (resp $y\_Tx,$ resp z_Tx) is the estimates of t_x when the initial estimates are computed with the x-plane (resp y-plane, resp z-plane)). Even if the initial estimates are very different, the results of the batch process with these 3 initializations are similar and the resulting curves are almost confounded.

Then we want to assess the accuracy of the recovered planes. Fig. 3.b shows the difference between the computed and the actual normal to the y-planes when the frames $1..i$, $(i \leq 40)$ are considered. Three curves are shown according to the plane chosen for the initial estimate. Once again, the normal to the planes are very similar whatever the considered initialization. The difference between the actual and the computed normal is around 4 degrees. Results on the third plane of the target are of the same order. Note that the camera moves very slowly at the beginning of the sequence. This explains why the results on the normal are not accurate when a small number of views are considered. It must also be noticed that the results using classical bundle adjustment techniques based on point correspondences leads to an error of the same order: if the 40 frame sequence is considered, the error on the normal is 4.29 degrees for the classical bundle adjustment using points and the error is 4.31 for our method. And our method is much more faster than the classical bundle adjustment.

Finally, we tested our algorithm on a set of sparse views. Using five equally spaced images in the considered sequence leads to an error on the normal vector of 4.91 degrees and 6.02 degrees on the two considered planes (the recovered equations for the 2 planes are: $.049x - y + .07z - 3.96 = 0; x - 0.686642y - 0.104819z - 6.234647 = 0$. The accuracy obviously depends on the choice of the set of images but the accuracy is quite satisfactory for scene augmentation.

To prove this, the recovered model of the target was used to perform real time registration. The viewpoint computed with this model is compared to the ones computed with the actual model in Fig. 1. A sphere was added to the scene in order to visually estimate the coherence of the augmented scene. Some snapshots of the augmented sequence are shown in fig. 4. Though we can notice that the sphere seems to slide slightly along the ground at some time instant, the visual impression is good and confirm that the accuracy of the model is sufficient for incrustation.

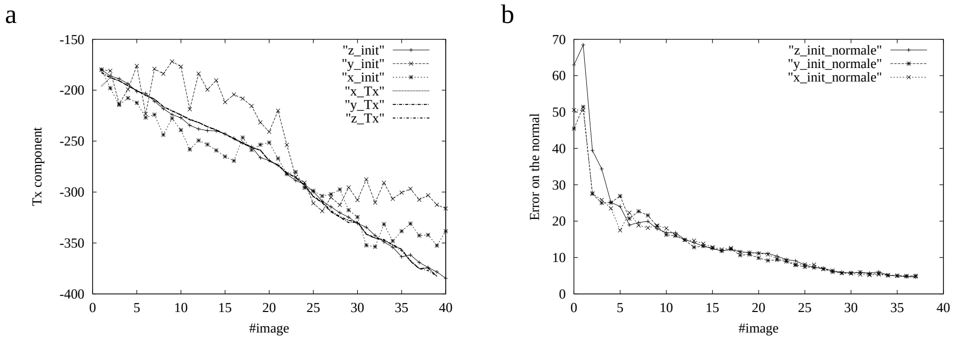


Figure 3: (a) The t_x component recovered by the structure and motion algorithm. (b) difference between the computed and the actual normal to the y-plane.

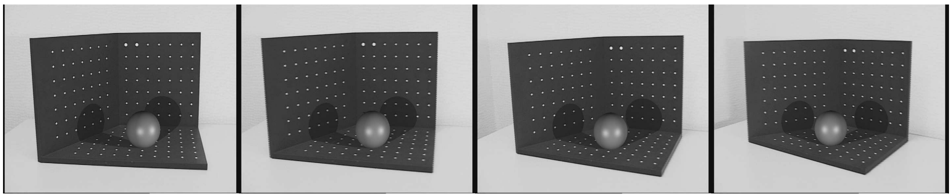


Figure 4: Some snapshots of the augmented scene using the recovered structure.

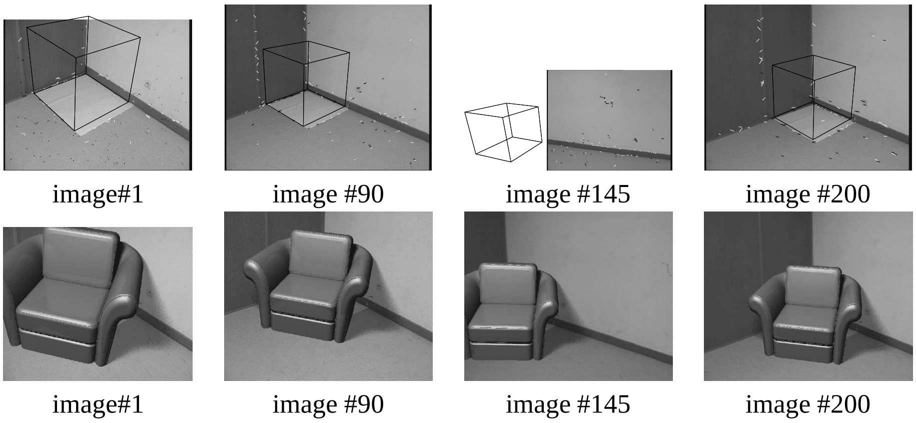
6.2 An indoor sequence

Results were also obtained on a 200-frames indoor sequence, shot in the basement of our laboratory. These experiments were conducted within a project of AR for e-commerce. One of the goal is to allow the user to visualize the products in its future environment. This sequence is particularly difficult to treat because the scene is very poorly textured (hardly a few stains on the ground and walls). Moreover, the camera motion is relatively fast in the second half of the sequence (up to 20 pixels of disparity between two images), and some images are blurred.

The projection matrix in first image was obtained using an ordinary poster laid on the ground. The aspect ratio was fixed to 1, and the principal point was assumed to lie at the center of the image. As a corner of the poster corresponded to the corner of the room, and because the angle between the two visible walls was a right angle, no measure had to be taken to know the equations of the three planes. Despite the difficulties mentioned above and the approximative knowledge of the internal parameters of the camera, the system succeeded in registering the two or three planes that were visible in the sequence. Fig. 5 shows the matching result and the projection of a cube after registration using LIN2, in four images of the sequence. Some snapshots of the scene augmented with a sofa are also shown in Fig. 5.

7 Discussion

A real time markerless registration system for augmented reality was presented. It provides accurate and reliable results for scenes which include planar structures. Our implementation yields results comparable in accuracy with full structure-and-motion methods but with better reliability and faster processing.



image#1 image #90 image #145 image #200

image#1 image #90 image #145 image #200

Figure 5: Firt Line: key-points matching and projection of a cube after registration in four images of the indoor sequence (black segments are inliers, white segments are outliers). Second line: some snapshots of the augmented scene.

Three methods were proposed for solving the n-planes registration problem. One of them (LIN2) has proven to be a good compromise between computation rates and accuracy of the composition. However, for long sequences, this method may progressively diverge because of successive approximations. This problem may be shaped by performing a bundle adjustment on a small number of images (the last five images for example, in the spirit of [5]). Of course, a hybrid system could also increase robustness and avoid drift by taking advantage of a partial 3D knowledge on the scene.

Finally, the possibility of recovering the multi-planar structure of the scene in a reasonable time and getting a good visual impression of the added scene from there is very promising. This must be investigated further for more complex real scenes.

Acknowledgements

This research is being supported with EC-grant IST-2000-28707 (ARIS project).

References

[1] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. Recent Advances in Augmented Reality. *IEEE Computer Graphics and Applications*, pages 34–47, 2001.

[2] A. Bartoli and P. Sturm. Projective Structure and Motion from Two Views of a Piecewise Planar Scene. In *Proc. International Conference on Computer Vision*, pages 593–598, 2001.

[3] D. Dementhon and L. Davis. Model Based Object Pose in 25 Lines of Code. *International Journal of Computer Vision*, 15:123–141, 1995.

[4] O. D. Faugeras and G. Toscani. The Calibration Problem for Stereo. In *CVPR 86, Miami, FL (USA)*, pages 15–20, 1986.

[5] A.W. Fitzgibbon and A. Zisserman. Automatic Camera Recovery for Closed or Open Images Sequences. In *ECCV'98, University of Freiburg (Germany)*, pages 311–326, June 1998.

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[7] P. Pritchett and A. Zisserman. Matching and reconstruction from widely separated views. In R. Koch and L. Van Gool, editors, *3D Structure from Multiple Images of Large-Scale Environments, LNCS 1506*, pages 78–92. Springer-Verlag, June 1998.

[8] G. Simon and M.-O. Berger. A Two-stage Robust Statistical Method for Temporal Registration from Features of Various Type. In *ICCV'98, Bombay (India)*, pages 261–266, January 1998.

[9] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In *Proc. International Symposium on Augmented Reality*, pages 120–128, 2000.

[10] G. Xu, J. Terai, and H. Shum. A Linear Algorithm for Camera Self-Calibration, Motion and Structure Recovery for Multi-Planar Scenes from Two Perspective Images. In *CVPR'2000, Hilton Head Island, South Carolina (USA)*, 2000.