Efficient initialization schemes for real-time 3D camera tracking using image sequences

Fakhr-Eddine Ababsa, Imane Zendjebil, Jean-Yves Didier, Malik Mallem

Efficient Initialization Schemes For Real-Time 3D Camera Tracking Using Image Sequences

Fakhreddine Ababsa, Imane Zendjebil, Jean-Yves Didier, M. Mallem
IBISC Laboratory, EA 4526, University of Evry
40, rue du Pelvoux, 91020 Evry Cedex - France
e-mail: {ababsa, zendjebil, didier, mallem}@iup.univ-evry.fr

Abstract— **3D camera tracking is an important issue for many kinds of applications such as augmented reality and robotics navigation. When image sequences are used, tracking needs to be initialized by matching 2D data extracted from images with 3D data representing a priori knowledge of the world. However, this process is difficult and the accuracy of the pose estimation strongly depends on the accuracy of the previous matching step. In this paper, we present two original approaches that achieve a 2D/3D points-based matching in order to initialize the 3D camera tracking. The first one is semi-automatic and requires the intervention of the user. The second one is completely automatic. Both approaches are based on SURF descriptors, which have the advantages of being fast and robust against outliers. A description of these two approaches is given and results obtained from experiments performed on real data are exposed and discussed.**

Keywords- Marker-less tracking, outdoor augmented reality, SURF descriptors.

I. INTRODUCTION

Augmented Reality systems (ARS) attempt to enhance human's perception of their indoors and outdoors working and living environments by complementing their senses with virtual input. Tracking computation refers to the problem of estimating the position and orientation (namely the pose) of the ARS user's viewpoint, assuming the user to carry a wearable camera [1][2]. Tracking computation is crucial in order to display the composed images properly and maintain correct registration of real and virtual worlds. Marker-less tracking provides an interesting issue because of its accuracy and robustness against the environmental influences (occlusions, illumination variation, etc.). The mean idea of the marker-less tracking is to identify, in the images, features from the 3D scene (points, lines, etc.).

However, a main problem with the marker-less tracking systems is the initialization process which provides the system automatically with the initial pose of the camera. This procedure needs to be done each time the system starts working or looses tracking. Several marker-less tracking approaches have been developed in the last years. In order to initialize their tracking system, Vachetti et al. [3] construct, during the off-line stage, a database using a set of keyframes, representing the scene from different viewpoints. Collected features consist on the two sets of corresponding 2D and 3D points and the camera projection matrix. The initialization is then done by matching features extracted from the initial image with the database using a similarity measure. The disadvantage of this method, beside its complexity, is that keyframes are local and thus sensitive to scale. Pressigout et al. [4] fuses a classical model-based approach based on edge extraction and a temporal matching relying on texture analysis into a single nonlinear objective function that has then to be minimized. Tracking is formulated in terms of a full scale non-linear optimization. The initialization is achieved by matching the first image with a set of multiscale reference images. This system works only with planar structures, and the initialization performance depends on the number of reference images. Genc et al. [5] propose a marker-less tracking based on learning framework of natural 3D features. However, their system needs an external marker based tracker to initialize the camera pose.

In this paper, we present two original and efficient approaches that achieve a 2D/3D points-based matching to initialize the 3D camera tracking for outdoor environments. We propose to use the SURF descriptors [6] in order to extract a stable set of natural key points which are invariant to image scaling and rotation, and partially invariant to changes in illumination and view point. We first propose a semi-automated initialization approach which needs the contribution of the user to perform reliable 2D/3D matching. In cases where tracking is lost, an automated initialization procedure is then launched.

The remainder of this paper is organized as follows. In section 2, we give the formulation of the camera pose estimation problem when using point features. Section 3 presents the overview of our proposed approaches. Obtained results on real data are then discussed in section 4.

II. CAMERA POSE PROBLEM FORMULATION

Throughout this paper, we assume a calibrated camera and a perspective projection model. If a point has coordinates $(x, y, z)^t$ in the coordinate frame of the camera, its projection onto the image plane is $(x/z\,,\, y/z, 1)^t$. In this section, we present the constraints for camera pose determination when using point and line features.

A. Problem definition

Let $\mathbf{p}_i = (x_i, y_i, z_i)^t, i = 1, \ldots, n,\ n \geq 3$ a set of 3D non-collinear reference points defined in the world reference

frame, the corresponding camera-space coordinates $\mathbf{q}_i = (x_i, y_i, z_i)$ are given by:

$$\mathbf{q}_i = R\mathbf{p}_i + T \quad (1)$$

where $R = (\mathbf{r}_1^t, \mathbf{r}_2^t, \mathbf{r}_3^t)^t$ and $T = (t_x, t_y, t_z)^t$ are a rotation matrix and a translation vector, respectively. R and T describe the rigid body transformation from the world coordinate system to the camera coordinate system and are precisely the parameters associated with the camera pose problem.

Let the image point $\mathbf{g}_i = (u_i, v_i, 1)^t$ be the projection of $\mathbf{p}_i$ on the normalized image plane. Using the camera pinhole model, the relationship between $\mathbf{g}_i$ and $\mathbf{p}_i$ is given by:

$$\mathbf{g}_i = \frac{1}{\mathbf{r}_3^t \mathbf{p}_i + t_z}(R\mathbf{p}_i + T) \quad (2)$$

which is known as the colinearity equation.

The point constraint corresponds to the image space error, it gives a relationship between 3D reference points, their corresponding 2D extracted image points and the camera pose parameters as follows:

$$E_i^p = \sqrt{\left(\hat{u}_i - \frac{\mathbf{r}_1^t \mathbf{p}_i + t_x}{\mathbf{r}_3^t \mathbf{p}_i + t_z}\right)^2 + \left(\hat{v}_i - \frac{\mathbf{r}_2^t \mathbf{p}_i + t_y}{\mathbf{r}_3^t \mathbf{p}_i + t_z}\right)^2} \quad (3)$$

where $\hat{\mathbf{m}}_i = (\hat{u}_i, \hat{v}_i, 1)^t$ are the observed image points.

The pose estimation problem is to find the rigid transform (R, t) that best fits the known 3D reference points with the observed 2D image points. Usually this is achieved by minimizing some form of accumulation of errors (least squares methods) based on equation 3. Typically Gauss-Newton or Levenberg-Marquardt methods are used for this purpose [7][8].

B. Discussion

3D-2D feature matching is critical for the camera pose estimation and still a difficult unsolved problem in computer vision. The tracking system needs an initialization that provides a set of good 3D-2D matched points. In practice, the accuracy of the matching process depends on the relevance of the information associated to the 3D points for their recognition. One interesting approach is to define a reference patches around the image points corresponding to the 3D model points. Matching is then performed by aligning these references patches within those extracted from the current frame. The correlation can be used to measure the similarity between the patches as in [9]. However, this method is not robust against illumination variation. Other approaches use the SIFT descriptors [10] for their robustness to changes in viewing conditions. The main disadvantage of the SIFT is its complexity and its high time consumption, this makes it not suitable for real-time applications.

In this work, we propose two initialization approaches. The first one is semi-automated and requires the user intervention to guide the matching process; it is used to start the tracking system. The second approach is fully automated; it is executed when the tracking is lost. The next sections give an overview of these approaches.

III. SEMI-AUTOMATED INITIALIZATION APPROACH

The proposed semi-automated initialization approach is performed in two steps: the wireframe model of the environment (here the building frontage) is first rendered, in real time, on the video flow coming from the camera, using a set of predefined poses (see figure 1-a). At the same time the user moves the camera in order to align the projected model within its image. Once this alignment is achieved (see fig. 1-b) the user validates the corresponding pose and the system switch to the matching step in order to perform, with high accuracy, the 3D-2D points matching. Aligning the rendered model allows to limit the search area of the 2D points in the current image. This makes the approach faster and robust against the outliers.

Figure 1. (a) the model environment rendering. (b) manual alignment between the projected model and its image

The 3D-2D matching is performed as follows. A search box is defined around each projected 3D model point on the aligned image. The interest points are then extracted from these image regions. As 3D points represent corners in the model, we use the Harris detector [11] in order to extract the 2D interest points. Then, a SURF descriptor is computed and associated to each extracted Harris point. We choose to use the SURF descriptor because it is scale-rotation invariant and allows real-time tracking. The distances between the reference descriptor associated to the 3D point and the descriptors of the extracted 2D points are measured and compared. The 2D point that minimizes this distance is selected as the corresponding 2D point. We have also used a RANSAC algorithm [12] in order to detect and remove outliers in the matching set, and thus increasing the accuracy of the initialization.

In order to validate the whole matching 3D-2D points, we introduce a coherence test which is used as a quality measurement for the estimated camera pose. We assume that this pose is close to that selected when the wireframe model is aligned within the image. Let, $P_a = [R_a \mid T_a]$ be the predefined camera pose that is used for the model/image alignment, and $P = [R \mid T]$ the camera pose estimated using the set of candidate matched points. As the two matrices are identical, we can write then:

$$P_a \cdot P^{-1} = I \tag{4}$$

Where I is a 4×4 identity matrix.
So, the trace of the matrix $P_a \cdot P^{-1}$ tends to 4. Our coherence test can be formulated as:

$$\delta < Trace(P_a \cdot P) < 4 \tag{5}$$

Where δ is a threshold below which the two matrices are considered different.

IV. Automated Initialization Approach

Unlike the semi-automated approach described above, in automated initialization procedure, the user intervention is not required. The system switches automatically in this mode every time when the tracking falls because of noise, occlusion or image blurring. This approach is performed as follows (see figure 2):

Let F_i be the reference frame that corresponds to the last captured image before the tracking failed. Much information are associated to this frame namely: the camera pose P_i and the set of matched 3D-2D points. The idea is to generate new 3D-2D matched points between the reference and the current frames.

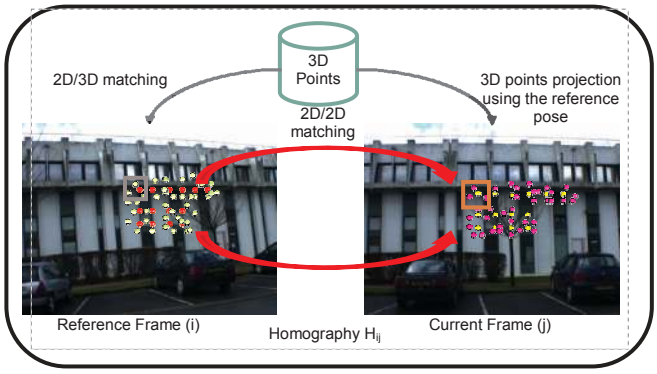


Figure 2. Automated initialization approach

For that, we first project the 3D points on the current frame using the reference camera pose Pi to generate predicted research areas. These image areas, named patches, are centered around the projected 3D points and have rectangular shape. The interest points corresponding to the SURF features are then extracted inside these patches and matched with those extracted from the reference frame. We also use a RANSAC algorithm in order to discard the outliers. To find the 3D-2D matched points for the current frame, we only need to identify the transformation that maps interest points defined in the reference image to those extracted in the current image. We assume that this transformation is a homography because the movement between the two frames is not meaningful. Let H_{ij} be this homography, m_i and m_j the interest points extracted from reference and current frames F_i, and F_j, respectively. The relationship between m_i, m_j and H_{ij} is defined as:

$$m_j = H_{ij} \cdot m_i \tag{6}$$

Once the homography is estimated, we apply it to the 2D image points associated to the 3D model points for the reference frame, in order to transform them in the current image. This allows updating correspondence between 3D and 2D points for the current image, and hence restarts automatically the tracking process.

V. Results

The proposed approaches have been tested in real scenes and the registration accuracy was analyzed. The results are presented for outdoor images (Figure 3) which corresponds to a moving camera pointing towards one frontage of a building. The frame rate of the recorded image sequences is about 25 frames/s and the resolution of the video images is 320×240 pixels. We used an industrial USB camera uEye UI-2220RE with a focal equal to 8mm.

The first experiment points out the performances of the semi-automated initialization approach in several conditions. Figure (3) shows that this approach performs good matching in spite of the illumination change.

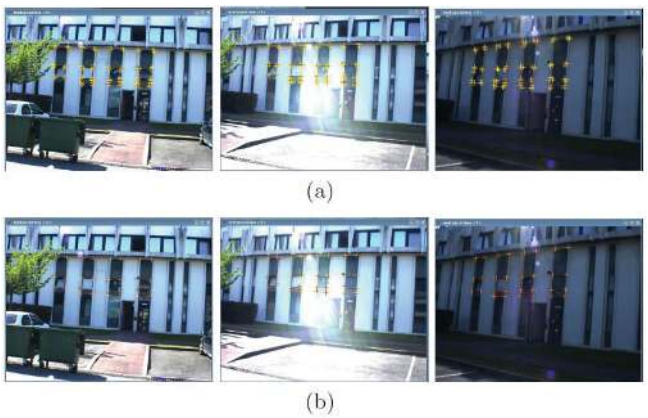


Figure 3.

Figure 4. Performance of semi-automated approach

Furthermore, in order to analyze the error in the matching process, we estimate the mean distance between the 2D points obtained after the semi-automated initialization step and the 2D points extracted from the images after refining the camera pose. We found a mean error equal to 3.8216 pixels with a standard deviation about 1.0873 pixels. This means that semi-automated approach is very efficient to generate a rough estimate of 2D-3D correspondences and really helps the tracking system to rapidly converge to the optimal solution.

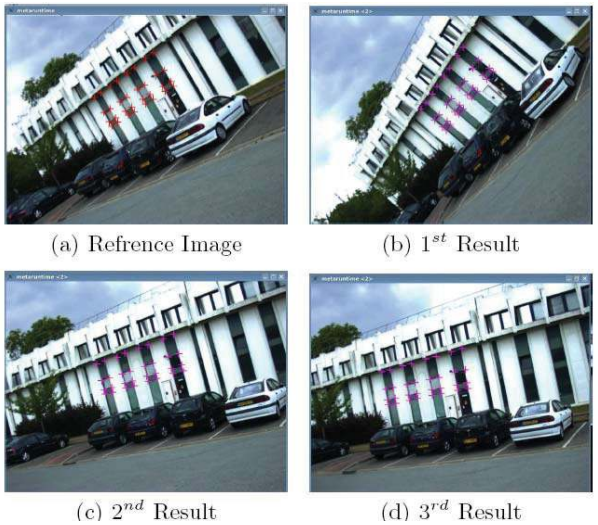
We also analyzed the execution time by carefully evaluating the processing time needed to achieve each step in the semi-automated initialization procedure. An example of these computation times is given in Table 1

TABLE I. COMPUTATION TIMES FOR THE SEMI-AUTOMATED INITIALIZATION APPROACH

Steps	**Times**
Manual alignment	unknown
Extraction and matching	50 ms
RANSAC	100 ms
Total (without manual alignment)	150 ms

This table shows that the computation time needed to achieve the semi-automated initialization matching is quite fast and makes this approach particularly efficient for the initialization stage of the tracking system.

In addition, we also tested the performances of our automated initialization approach. For that we have considered several images taken under different viewpoints. Figure 4 shows some obtained results. We can see that, for all the considered cases, the reference points (taken on the frontage of the building) are well matched in the current frames. In this example we have considered coplanar points.

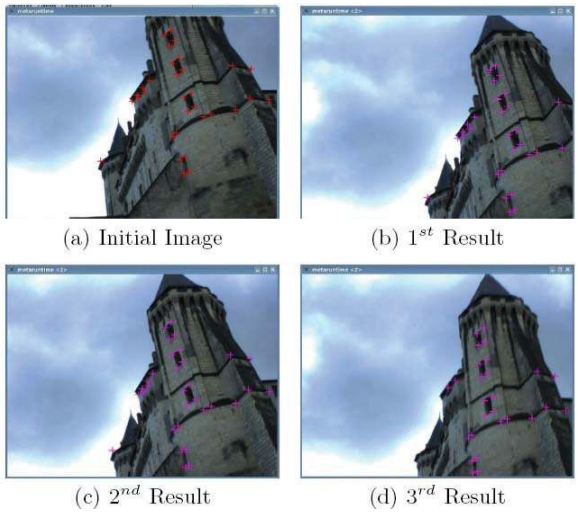


(a) Refrence Image (b) 1^{st} Result

(c) 2^{nd} Result (d) 3^{rd} Result

Figure 5. Matching results for autometed initialization approach, case of the coplanr points

We have also tested our approach for non coplanar points chosen on the tower of the castle (figure 5). Obtained results in this case are satisfying. Indeed, combining the SURF points with the RANSAC algorithm provides an accurate and robust homography estimation, and thereby allows good points matching. The matching process in this case gives a mean error about 1.7823 pixels with a standard deviation of 0.6634 pixels.



(a) Initial Image (b) 1^{st} Result

(c) 2^{nd} Result (d) 3^{rd} Result

Figure 6. Matching results for autometed initialization approach – Non coplanr points

The computation time of this approach depends mainly on the SURF features extraction and matching. Introducing a prediction stage in order to limit the research area of the interest points in the current frame has significantly reduced the total time of the whole algorithm (practically, it is divided by two). Comparing to other similar approaches [7], our proposed automated initialization technique is more flexible and provides best real times performances.

VI. CONCLUSION

We have presented, in this paper, two initialization approaches for marker-less visual tracking, namely a semi-automated and fully automated 3D-2D points matching. Both approaches use the SURF descriptors in order to achieve the matching robustness against the illumination changes and rotations. The semi-automated approach requires the user intervention in order to perform model/image registration when the tracking starts, while the automated approach is executed every time when the tracking falls. Our main idea consists in associating 2D patches composed of the SURF features to the 3D model points, allowing their efficient recognition in the image sequence. Both approaches have been tested on real data and have demonstrated satisfying results. However, some improvements should be done to improve more the efficiency of our tracking system. Future research efforts will include the use of other kind of sensors, like inertial and GPS, in order to assist the visual tracking and improve its accuracy.

ACKNOWLEDGMENT

This work is supported by the RAXENV project funded by the French National Research Agency ”ANR”.

REFERENCES

[1] F. Ababsa, M. Mallem,”Robust line tracking using a particle filter for camera pose estimation. Proc. Virtual Reality Software and Technology (VRST’06), pp. 207-211, 2006.

[2] F. Ababsa, M. Mallem: A Robust Circular Fiducial Detection Technique and Real-Time 3D Camera Tracking. Journal of Multimedia , vol. 3, n°4, pp.: 34-41, 2008.

[3] L. Vacchetti, V. Lepetit, and P. Fua, “Stable real-time 3d tracking using online and offline information”. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 26, n°10, pp. 1385-1391, 2004.

[4] M. Pressigout, and E. Marchand, “Real-Time Hybrid Tracking using Edge and Texture Information”. Int. Journal of Robotics Research, IJRR, vol. 26, n°7, pp. 689-713, 2007

[5] Y. Genc, S. Riedel, F. Souvannavong, C. Akinlar, and N. Navab. “Marker-less tracking for ar: A learning-based approach”. Proc. Int. Symp. Mixed and Augmented Reality (ISMAR’02), pp. 295-304, October 2002.

[6] H. Bay, A. Ess, T. Tuytelaars, and L.Van Goo, “SURF: Speeded Up Robust Features”. Computer Vision and Image Understanding (CVIU), vol. 110, n°3, pp.346-359, 2008.

[7] D.G. Lowe, “Fitting Parameterized Three-Dimensional Models to Images”. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 13, pp. 441-450, 1991

[8] R.M Haralick, “Pose Estimation From Corresponding Point Data”. IEEE Trans. Systems, Man, and Cybernetics, vol. 19, n°6, pp.1426-1446, 1989.

[9] G. Bleser, D. Stricker, “Advanced tracking through efficient image processing and visual-inertial sensor fusion”. Computer & Graphics, Elsevier, New York, February 2009

[10] D.G Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. Int. J. Comput. Vision, vol. 60, n° 2, pp. 91-110, 2004.

[11] C. Harris. “Tracking with rigid models, Active vision”, Cambridge, MA, USA, MIT Press, pp. 59-73, 1193.

[12] M. A. Fischler, R. C Bolles. “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”, Commun. ACM, New York, NY, USA, vol. 24, n°6, pp. 381-395, 1981.