In Proceeding of WIAMIS 2005, Montreux, Switzerland

ONLINE REGISTRATION TOOL AND MARKERLESS TRACKING FOR AUGMENTED REALITY

David Marimon i Sanjuan, Yousri Abdeljaoued and Touradj Ebrahimi

Signal Processing Institute (ITS)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{david.marimon, yousri.abdeljaoued, touradj.ebrahimi}@epfl.ch

ABSTRACT

This paper presents a markerless tracking algorithm for augmented reality with no constraints in the geometry of the scene. User's pose is estimated using natural feature point extraction and tracking, and the epipolar geometry constraint between camera views. An online registration tool is also described. With this procedure, the user determines the initial 3D pose of the virtual object that augments the scene. Validation and discussions about advantages of this algorithm along with further steps to be taken are presented.

1. INTRODUCTION

Merging real and virtual worlds is a growing research area since the 1990s. Augmented Reality (AR) is the overlay of information (objects, text, etc.) on a real scene. This additional information is registered with the real environment seamlessly. AR systems perform this task in real time and, in some cases, actively interacting with the user.
AR interfaces in several areas such as medical surgery, design, entertainment and machinery assembly have shown interesting improvements [1].
Research in AR pays a particular attention to the problem of tracking. The estimation of the 6DOF (position and orientation) of the user allows the inclusion of the virtual object in the real scene. Among the available technologies to track position and orientation are acoustic, optical, mechanical, magnetic, inertial and hybrid trackers [12]. A subset of optical tracking called video tracking is of special interest. Among video-based tracking techniques, markerless video tracking analyzes the motion of natural features. The techniques presented in [8, 11] use natural features in planar structures in the scene to calculate transformations between frames. Other techniques use the epipolar constraint theory. These determine the projection between views using point correspondence with no geometrical limitation [3, 5].
Several problems have been identified in AR environments. We address in this paper the problem of scene preparation and initial registration of the scene. Techniques such as marker-based and model-based tracking have proven impressive tracking results but they still rely on scene preparation. Marker-based tracking uses fiducial markers where virtual objects can be registered. Model-based tracking uses a CAD model of the object in the scene that is augmented. Markerless tracking deals with the limitations of these tracking techniques. Tracking of natural features avoids any preparation. Another major contribution of markerless tracking is its larger functional range. Initial registration of the scene in markerless tracking techniques such as [3] still relies on markers. Our purpose is to have fully markerless AR. User's capability to register a scene (Spatial Ability) using virtual cursors has proven good results even with a single camera set [10].
The technique proposed in this paper uses natural features to estimates user's pose based on the epipolar constraint. This ensures no limitation on the geometry of the scene. In addition, an online registration tool is described. This initialization step gives the user the opportunity to place a virtual object wherever in the scene.
The structure of this paper is as follows. The technique is presented in Section 2, including the steps taken for correct track update and pose estimation, together with a description of the online registration tool. Performance test are dealt with in Section 3. Finally, we discuss the advantages and disadvantages of the proposed system in Section 4.

2. PROPOSED TRACKING TECHNIQUE

This work presents a system to perform markerless tracking for AR using natural features in the scene. This system gives a framework for AR in unprepared environments relying on the performance of Feature Point (FP) tracking. In order to accurately register the scene, user's pose has to be recovered. Two steps are to be taken: FP tracking and pose estimation from these tracks. The novelty in the proposed approach is its online registration tool. The user can manu-

ally chose the initial 3D position of the virtual object added to the real environment. Figure 1 shows the block diagram of the system. First, the user is asked to select the desired

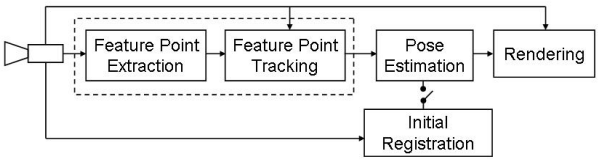


Fig. 1. Block diagram of the system.

3D position of the virtual object. Once this initialization is completed, video frames are processed for FP extraction. Tracks are updated with this extracted FPs. For every new frame, the camera projection matrix is calculated using the tracks. After the projection is obtained, the scene can be augmented.

2.1. Feature Point Extraction and Tracking

The system's performance relies heavily on FP tracking quality. Motion information extracted from video stream depends on tracks' stability. Natural points, referred to also as FP, have demonstrated easy detectability and rather slow appearance mutation.

The procedure is performed on a frame-by-frame basis. For each new frame, several FPs are extracted and all tracks are updated. An intensity-based algorithm is applied for point extraction. Every track is matched against all the extracted candidates to find its new position. Robust association is fundamental to ensure good quality of tracks. Two sequential selection procedures are followed: (i) candidates closer than a certain Mahalanobis distance from the track coordinates are selected; (ii) an intensity test is performed with the remaining candidates. The distance is calculated with the innovation covariance of a Kalman filter associated to each track. By doing so, candidates inside track tendency are favored. For the intensity test, previous gray-level pixel neighborhood is correlated with that of the candidate . If no match is found for a track, it is terminated. This can be caused by a FP escaping the view, being occluded or simply not extracted. Extracted FPs that have not been associated with any existing track are considered as new tracks. Tracking user's pose has to deal with maneuvering stages. An Interacting Multiple Model estimator [2] is used to switch between two models: namely a small maneuvering model and a fast maneuvering model. Both implemented with a Kalman Filter.

2.2. Pose Estimation

Pose estimation of the user permits a correct augmentation of the scene. Since in our case the camera is attached to the user's head, the rotation (R) and translation (t) of the camera represent the motion of the user. Pose estimation is resolved by calculating the camera projection matrix

$$P = K[R|t]$$

in real time. Where K is the calibration matrix calculated offline with the method proposed in [6].

2.2.1. Camera Projection

The system analyzes the motion of the scene using natural features in view. Correspondence of FPs between views can be used to extract rotations and translations of the camera. Epipolar geometry defines the projection between two views. It depends only on the camera's internal parameters and is independent from the scene's structure. At the core of this theory is the Fundamental Matrix (F) that sustains the above mentioned correspondences between views [7]. A point $\mathbf{X}$ is imaged in two views of the same scene as $\mathbf{x}$ and $\mathbf{x}'$ with camera projection matrices $P = K[I|0]$ and $P' = K[R|t]$, respectively, where I is the identity matrix. The relation of both imaged points is given by the Fundamental Matrix in (1).

$$x'^T F x = 0 \tag{1}$$

If cameras are calibrated, one can use the Essential Matrix (E) instead. This is a particularization of F to the case of normalized image coordinates, in other words, compensating the coordinates distortion with the calibration. Resolving F using (1) leads to a projective ambiguity. However, resolving $x'^T E x = 0$ leads to an ambiguity of scale. Generally, there are four solutions to this equation but only one is geometrically possible in our case, namely, that corresponding to both views in front of the camera. In this case, P' can be calculated directly as follows: suppose that the Singular Value Decomposition (SVD) of E is $U diag(1,1,0) V^T$, then, if the first camera matrix is $P_{norm} = [I|0]$ (using normalized coordinates), the second camera matrix is [5]:

$$P'_{norm} = \left[U \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} V^T \,\middle|\, +u_3 \right] \tag{2}$$

where U and V are the orthogonal matrices extracted by the SVD, and u_3 is the third column of U. Details about how to compute the SVD can be found in [9].

Among the possibilities to solve the Essential Matrix equation, we have chosen the 7-point algorithm (a set of seven point correspondences to solve Equation 1). Current projection P_k is calculated from the geometrical relationship between two frames $[R_k|t_k]$ and the previous projection $P_{k-1} = [R_{k-1}|t_{k-1}]$ as follows:

$$P_k = [R_k \cdot R_{k-1} | t_k + t_{k-1}] \tag{3}$$

The update of some tracks in FP tracking step may be incorrect. Consequently, some correspondences between feature points in the previous and the current frame are false. In order to reduce their effect on the estimation process, robust estimation techniques are needed. The iterative RANSAC algorithm [4] is applied to deal with outliers in the set.

2.3. Online Registration Tool

The goal here is to give the initial coordinates of a virtual object (VO). The system does not reconstruct the 3D scene, but rather exploits the Spatial Ability of a human being to register it. The user assists the system such that it can determine the desired 3D position of the VO. This is achieved by manually selecting VO's position in space. The user perceives the reality in front of him through a small camera connected to a video see-through head-mounted display. Depth and size of VO is calculated with the following assumption. First, there is no rotation and no translation since the user will place the VO in known coordinates. The algorithm presented before for pose estimation assumes that the previous projection matrix has neither rotation nor translation (See Equation 2). From a projection point of view, a point $\mathbf{X} = (X, Y, Z)^T$ is mapped to the image plane in $\mathbf{x} = (x_{cam}, y_{cam})^T$. In our case, we will use a CCD camera where pixels are not exactly square (α_x, α_y) and the origin of coordinates has an offset from the principal point (x_0, y_0). All these parameters are obtained with camera's off-line calibration. The projection matrix at this step is $P = K[I|0]$. Equation 4 shows the projection between 3D and 2D.

$$\begin{pmatrix} Z \cdot x_{cam} \\ Z \cdot y_{cam} \\ Z \end{pmatrix} = \begin{pmatrix} \alpha_x & 0 & x_0 & 0 \\ 0 & \alpha_y & y_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \quad (4)$$

Where x_{cam} and y_{cam} are the pixel coordinates in the camera image plane.

The online registration requires that the user, using the pointer of a mouse, selects a 2D region in the image plane. This proportionates an initial 2D position and the depth of the VO. In a first step, the user is asked to draw a 2D square of surface $\Delta X \cdot \Delta Y$ (e.g. 1cm x 1cm of the real world) with sides parallel to $\mathbf{X}$ and $\mathbf{Y}$ axes. This square is drawn in the image plane so actually the user is defining a square of size $\Delta x_{cam} \cdot \Delta y_{cam}$ pixels^2. Figure 2 illustrates the geometric representation of this VO initialization process. In a second step, the depth (Z_{VO}), assuming the real square is parallel to the image plane, can be calculated as follows:

$$Z_{VO} = \alpha_x \frac{\Delta X}{\Delta x_{cam}} = \alpha_y \frac{\Delta Y}{\Delta y_{cam}} \quad (5)$$

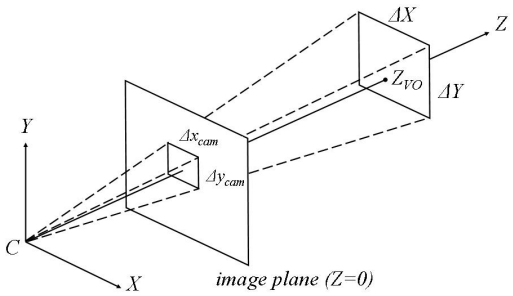


Fig. 2. Geometric representation of virtual object's position initialization. C is the camera center.

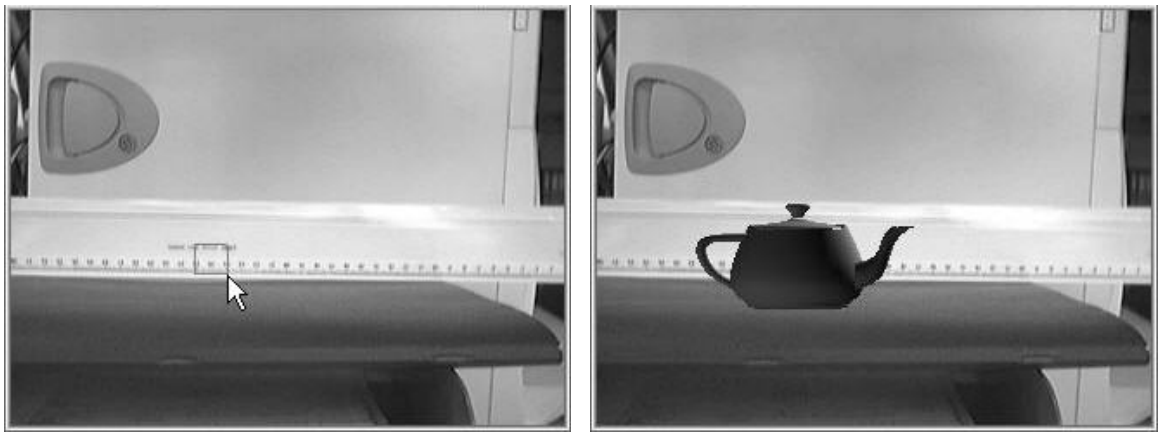
This depth is extracted from a single camera view right at the time when the user ends drawing the square. The real world square unit (in the previous example cm^2) is put in correspondence with the virtual size unit of the VO's 3D model. The VO is then resized accordingly and inserted in the video frame, augmenting the scene. We maintain the coordinate system of the VO. One of the vertices of the 2D region drawn, corresponds to the origin of coordinates for the VO. The algorithm presented in the previous section determines the following poses of the VO.

3. EXPERIMENTS

This section presents validation and performance tests of the feature point extraction and tracking technique, and the online registration tool.

The tests on the extraction and tracking of feature points shows that the IMM framework tracks both non-maneuvering and maneuvering phases. We obtain a Root Mean Square Error (RMSE) of 1.97 pixels in a sequence of 60 frames with several direction changes. We compared this result with a single Kalman Filter framework. Once a maneuver starts, this filter loses the track.

Following is a description of a real online registration case. Figure 3 shows the procedure needed to register the scene. In this case, the user was assisted using a real 50 cm ruler situated 49 cm away from the camera and with its axis perpendicular to that of the camera. The user was asked to select a real world 2cm x 2cm square (Figure 3(a)) and then the VO was placed accordingly (Figure 3(b)). In this case the system calculated two different measures for Z_{VO} (See Equation 5). The depth estimated given Δx_{cam}, was 48.06 cm. Given Δy_{cam}, 51.02 cm. The reasons for such deviation are the distance from the camera, resolution of the image (affects alignment ability), and lack of ruled assistance in y coordinates. In cases where neither alignment nor depth precision is crucial, biometrics such as the length of a hand's phalange, for example, could be used as a distance



(a) The user draws the square in the image plane (real 2cm x 2cm). (b) The VO is sized and placed accordingly.

Fig. 3. Online Registration using a ruler.

reference. In addition, for this example, the augmented reality would be particularly adapted to the user.

4. DISCUSSION

This paper presents the ongoing development of a system for augmented reality in unprepared environments. Additionally, it assumes no constraint on the geometry of the scene. Such a system could be applied to any environment even when its content changes in time, by repeating the initialization step. The main advantage of the technique presented here is that the user can freely control the appearance of his surroundings anytime during his experience.

As a matter of fact, the quality of FP extraction and tracking is the key for subsequent steps. An erroneous extraction can lead to a bad tracking. Consequently, this will affect the successive positions and orientations of the VO as the projection depends on previous calculations (see Equation 3). Techniques that recover from false projections should be added. In this direction, degeneracies such as homography case, should be addressed. A different problem arises from the online registration tool. If a precise scaling is needed (such as in augmented medical surgery), the spatial ability of the user can be a determining factor. In this case, computer-assisted registration would be a possible enhancement.

The proposed algorithm brings the advantages of a robust FP tracking for camera motion estimation to an augmented reality framework. Online registration introduces a fast tool for augmented framework's set up. The less cumbersome the set up of an environment, the more flexible it becomes to changes.

5. REFERENCES

[1] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. Recent advances in augmented reality. *IEEE Computer Graphics and Application*, 21(6):34–47, Nov 2001.

[2] Y. Bar-Shalom and W. D. Blair. *Multitarget-Multisensor Tracking Applications and Advances - Volume III*, chapter 3. Artech House, 2000.

[3] K. W. Chia, A. Cheok, and S. Prince. Online 6 dof augmented reality registration from natural features. In *Proc. of the ISMAR*, pages 305–313, Sep–Oct 2002.

[4] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. In *Communications of the ACM*, volume 24, pages 381–395, 1981.

[5] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[6] H. Kato and M. Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Proc. of the 2nd IEEE and ACM IWAR*, pages 85–94, Oct 1999.

[7] Q.-T. Luong and O. Faugeras. The fundamental matrix: theory, algorithms, and stability analysis. In *Intl. Journal of Computer Vision*, volume 17, pages 43–76. 1996.

[8] U. Neumann and S. You. Natural feature tracking for augmented reality. *IEEE Transactions on Multimedia*, 1(1):53–64, Mar 1999.

[9] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[10] J. Sands, S. Lawson, and D. Benyon. Do we need stereoscopic displays for 3d augmented reality target selection tasks? In *Proc. 8th Intl. Conf. on Information Visualization (IV'04)*, pages 633–638, 2004.

[11] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In *Proc. IEEE and ACM ISAR*, pages 120–128, Oct 2000.

[12] G. Welch and E. Foxlin. Motion tracking: no silver bullet, but a respectable arsenal. *Computer Graphics and Applications*, 22(6):24–38, Nov–Dec 2002.