Computing and Informatics, Vol. 27, 2008, 891–911

ROBUST ESTIMATION OF TRIFOCAL TENSORS USING NATURAL FEATURES FOR AUGMENTED REALITY SYSTEMS

Tao Guan*, Lijun Li, Cheng Wang

Digital Engineering and Simulation Centre
HuaZhong University of Science and Technology
430074, Wuhan, China
e-mail: qd_gt@126.com

Manuscript received 7 September 2006; revised 26 September 2008
Communicated by Dieter Kranzemüller

Abstract. Augmented reality deals with the problem of dynamically augmenting or enhancing the real world with computer generated virtual scenes. Registration is one of the most pivotal problems currently limiting AR applications. In this paper, a novel registration method using natural features based on online estimation of trifocal tensors is proposed. This method consists of two stages: offline initialization and online registration. Initialization involves specifying four points in two reference images respectively to build the world coordinate system on which a virtual object will be augmented. In online registration, the natural feature correspondences detected from the reference views are tracked in the current frame to build the feature triples. Then these triples are used to estimate the corresponding trifocal tensors in the image sequence by which the four specified points are transferred to compute the registration matrix for augmentation. The estimated registration matrix will be used as an initial estimate for a nonlinear optimization method that minimizes the actual residual errors based on the Levenberg-Marquardt (LM) minimization method, thus making the results more robust and stable. This paper also proposes a robust method for estimating the trifocal tensors, where a modified RANSAC algorithm is used to remove outliers. Compared with standard RANSAC, our method can significantly reduce computation complexity, while overcoming the disturbance of mismatches. Some experiments have been carried out to demonstrate the validity of the proposed approach.

[*] corresponding author

Keywords: Augmented reality, registration, trifocal tensors, projective reconstruction

1 INTRODUCTION

Augmented Reality (AR) copes with the problem of dynamically augmenting or enhancing the real world with computer generated virtual objects. Unlike virtual reality, AR does not create a simulation of reality. Instead, it takes a real object as the foundation and incorporates technologies that add contextual data to deepen a person's understanding of the subject. Augmented Reality has been put to use in a number of fields, including medical imaging [1], where doctors can access data about patients; training [2], in which technology provides students or technicians with necessary data about specific objects they are working with; and in museums [3], where artifacts can be tagged with information such as the artifact's historical context or where it was discovered.

Registration is one of the most pivotal problems currently limiting AR applications. It means that the virtual scenes generated by computers must be aligned with the real world seamlessly. Current registration methods can be divided into two categories: sensor based registration and computer vision based registration [4].

Registration methods based on sensors such as magnetic [5] or ultrasonic [6] are feasible under different circumstances, including illumination change and rapid movement of user's head. But these kinds of methods suffered from the following disadvantages: First, such devices are generally too expensive for common consumers to buy. Second, the bulk and weight of them are also inconvenient for users to mount. Third, the position and the orientation information generated by sensors are usually not precise enough to achieve seamless registration.

Computer vision based registration can solve the above problems to some degree. These kinds of methods do not require any special and expensive equipment except for common cameras and a personal computer. The vision-based approaches can also be divided into marker-based approach and natural feature-based approach.

Marker-based approach is the most commonly used registration method in the majority of applied AR systems. In these systems, one or more man-made fiducial markers are put into the scenes, and the camera's position and orientation are computed using these markers' projections on the moving frame. The downfall of this approach is that the fiducial markers must be within the area of user's view from beginning to end. Otherwise, the registration will be invalidated.

To overcome problems of the man-made markers, numerous natural features based registration approaches have been put forward. These kinds of method take full advantage of natural features including points [7, 8, 9], planes [10, 11, 12], curves [13, 14] and so on, to achieve registration between real and virtual scenes. By using natural features, the user's movement range is not limited, and the original scenery is not broken. Therefore natural features based registration has become one of the most active research topics in AR field.

Previous work on markerless camera tracking can be coarsely classified into the following groups:

- Model-based approaches like [15, 16, 17] require a complete 3D model of the target scene and often need a large amount of pre-processing. This reduces online processing time and makes the tracking quite stable. However, constructing a 3D model of target environment or objects and obtaining templates around feature points are troublesome tasks which require much effort. Moreover, these kinds of methods restrict the camera motion to the modeled environment and maintenance is complex, if the scene changes.
- Classical structure from motion (SFM) approaches, which start from scratch and recover scene structure and camera motion simultaneously, are known to suffer from drift. Simon et al. [10, 11] proposed a registration method using planar structures in the scenes. They calculated the projection matrix by real-time estimation of homographies between consecutive frames. Whereas this method suffered from the problem of error accumulation, a reference plane must be specified and other planes need to be perpendicular to this plane under the condition of multiple planes. Yuan et al. [8, 9] proposed a method by which a user can specify four points, which form an approximate square, to define the world coordinate system. The projections of these four specified points are computed in the live video sequence and used to estimate the camera pose. In this method, the projective reconstruction technique is used to set up an updated projective transformation between the image points and the 3D projective space. Experiments showed that this method is unstable when the camera moves. Pang et al. [7] made use of affine reconstruction and reprojection techniques to estimate the image projections of the four specified points used to establish the world coordinate system in the live video sequence. These image projections were then used to estimate the camera pose in real time. However, this method does not consider tracking the feature points robustly and is prone to being disturbed by mismatches. Moreover, special square planar structure is still used in initializing stage to construct world coordinate system. Nister et al. [18] had reported on first success with real-time pose estimation in completely unknown scenes. However, long-term stability is poor (the algorithms tend to drift), and drift is limited by insertion of a "firewall", which leads to a system restart when the drift becomes too large [19].
- A different method for simultaneous structure and pose recovery is the SLAM approach (Simultaneous Localization and Mapping). Much work has been performed in this area [20, 21, 22]. The system developed by Davison provides two crucial points for obtaining good tracking performance and minimal drift: the tracking of strongly salient features, which function as long term landmarks, and the consequent propagation of uncertainty in their locations. This is done by maintaining a full correlation between the camera and all features in the scene,

which makes insertion and deletion of features not trivial and for reasons of efficiency limits the total number of features that can be tracked at a time.

In this paper, we follow the structure from motion approaches and propose a novel markerless registration method based on robust estimation of trifocal tensors using natural features. Our method distinguishes itself in following three ways:

1. Motivated by the principle that the homography can be calculated using the correspondence of four noncollinear points, we relax the restriction that the four specified points used to establish the world coordinate system must form an approximate square. The only limitation of our approach is that these four coplanar points should not be collinear.
2. Benefiting from the tensor of previous frame and the normalized cross-correlation (NCC) operation, we propose a new method to match features between current and reference images directly. By this method, not only do we overcome the problem of losing features, but also constitute a NCC based criterion to evaluate the quality of point matches, which is a very important necessity of the method we used to calculate the needed tensor.
3. To estimate trifocal tensors precisely, we propose a modified RANSAC algorithm to remove outliers. The matches with higher quality (normalized cross-correlation score) are tested prior to the others, by which the algorithm can arrive at termination criterion and stop sampling earlier. Compared with the standard RANSAC algorithm, our method is more stable and can reduce sample times to a large degree, which enables our system suitable for online implementation.

The rest of this paper is organized as follows. Section 2 presents some background and notations. A primitive overview of the proposed method is described in Section 3. Section 4 presents our new method to set up the world coordinate system. Section 5 deals with the problem of feature tracking. Section 6 discusses the modified RANSAC based tensor estimation method. Section 7 gives the nonlinear least square based pose optimization method. Experimental results are shown in Section 8. Final remarks and conclusion are given in the last section.

2 PRELIMINARIES

2.1 Camera Model and Homography

In our research, both 2D and 3D points are represented by homogeneous vectors, so that the relationship between a 3D point $X = (X, Y, Z, 1)^T$ and its image projection $x = (x, y, 1)^T$ is generally given as follows under the pinhole camera model

$$x = \lambda PX \text{ where } P = K[R|T] \tag{1}$$

where λ is the homogeneous scale factors unknown a-priori, P is the projection matrix, $[R|T]$ is the extrinsic parameter of the camera, which is also called registration matrix in our paper, $R = [r_x \; r_y \; r_z]$ is the 3×3 rotation matrix, and $T = [t]$ is the translation of the camera. The matrix K represents the intrinsic parameters of the camera:

$$K = \begin{bmatrix} f & s & u_0 \\ 0 & af & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where f is the focal length, (u_0, v_0) are the coordinates of the principal point, a is the aspect ratio, s is the skew factor of the image axes. In our method, we assume that the intrinsic parameters are known in advance and do not change, as it is reasonable in most cases.

When a 3D point exists on the $Z = 0$ plane of the world coordinate, Equation (1) will be [10]

$$\begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = \lambda K \left[r_x^i r_y^i r_z^i t^i \right] \begin{pmatrix} X_w \\ Y_w \\ 0 \\ 1 \end{pmatrix} = \lambda K \left[r_x^i r_y^i t^i \right] \begin{pmatrix} X_w \\ Y_w \\ 1 \end{pmatrix} = H_w^i \begin{pmatrix} X_w \\ Y_w \\ 1 \end{pmatrix} \tag{2}$$

where the 3×3 matrix H_w^i is a planar homography which transforms points on the world plane to the current i^{th} image. Given four or more point correspondences, we can compute H_w^i using singular value decomposition. Then registration matrix $[R^i|T^i]$ can be recovered as follows:

$$\left[R^i|T^i \right] = \left[r_x^i r_y^i (r_x^i \times r_y^i) t^i \right] \tag{3}$$

2.2 Fundamental Matrix and Projective Reconstruction

The epipolar geometry exists between any two-camera systems. For a point x_i in the first image, its correspondence in the second image, x_i', must lie on the epipolar line in the second image, which is known as the epipolar constraint. Algebraically, in order for x_i and x_i' to be matched, the following equation must be satisfied [23]:

$$x_i' F x_i = 0 \quad i = 1, \ldots, n \tag{4}$$

where F, known as the fundamental matrix, is a 3×3 matrix of rank 2, defined up to a scale factor, which is also called an essential matrix in the case of two calibrated images. In our research, although the intrinsic parameters are known in advance, the fundamental matrix is still calculated in order to construct the camera projective matrices from this fundamental matrix.

Let F be the fundamental matrix between two images. It can be factored as a product of an antisymmetric matrix $[e']_x$ and a matrix T, i.e., $F = [e']_x T$. In fact,

e' is the epipole in the second image. Then, two projective camera matrices can be represented as follows:

$$P = [I|0], P' = [T|e'] \tag{5}$$

Given a pair of matched points in two images: (x_i, x_i'), let X_i be the corresponding 3D point. From Equation (1), the following two equations can be obtained:

$$x_i = \lambda P X_i \tag{6}$$

$$x_i' = \lambda' P' X_i \tag{7}$$

where λ and λ' are two arbitrary scalars. Let p_i and p_i' be the vectors corresponding to the i^{th} row of P and P', respectively. The above two scalars can be computed as follows:

$$\lambda = 1/p_3^T X_i \tag{8}$$

$$\lambda' = 1/{p_3'}^T X_i \tag{9}$$

With Equations (6)–(9), we can reconstruct X_i from its image matches (x_i, x_i') using the linear least square technique. Details about projective reconstruction can be found in [9] and [23].

2.3 Trifocal Tensor

The trifocal tensor plays a similar role in three views compared to that played by the fundamental matrix in two. It encapsulates all the projective geometric relations between three views that are independent of scene structure [24]. For a triplet of images, the image of a 3D point X is x, x' and x'' in the first, second and third images respectively, where $x = (x_1, x_2, x_3)^T$ are homogeneous three vectors. If the three camera matrices are in canonical form, where $P = [I|0]$, $P' = [a_j^i]$, $P'' = [b_j^i]$, and the a_j^i and b_j^i denote the ij^{th} entry of the matrix P' and P'' respectively, index i being the row index and j being the column index, then the trifocal tensor can be computed by

$$T_i^{jk} = a_i^j b_4^k - a_4^j b_i^k, \quad j,k = 1,\dots,3, i = 1,\dots,3 \tag{10}$$

where $T = [T_1, T_2, T_3]^T$ is a $3 \times 3 \times 3$ homogeneous tensor. Using the tensor, a point can be transferred to a third image from correspondences in the first and second images:

$$X_l'' = x_i' \sum_{k=1}^{3} x_k T_k^{jl} - x_j' \sum_{k=1}^{3} x_k T_k^{il}, \quad i,j = 1,\dots,3, l = 1,\dots,3 \tag{11}$$

The details for calculating trifocal tensor will be discussed in Section 6.

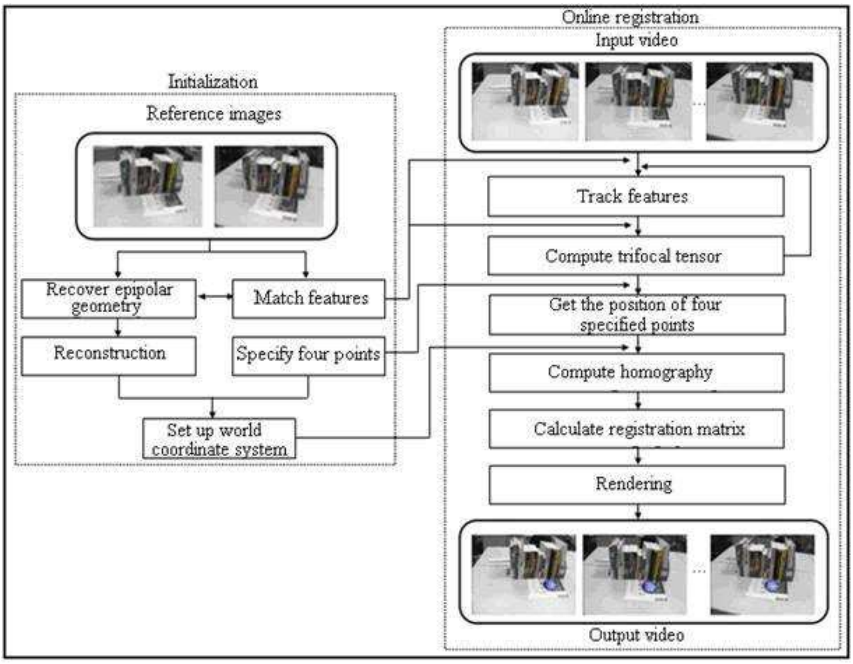


Fig. 1. Outline of the proposed approach

3 OVERVIEW OF PROPOSED APPROACH

This section explains the algorithm for our registration method based on robust estimation of trifocal tensors using natural features. It can be divided into two stages, namely offline initialization and online registration as shown in Figure 1.

In the initialization stage, two spatially separated images of the scenes in which we want to augment are selected as the references. A high quality set of point feature [25] matches and the fundamental matrix are obtained for these two images by the normalized cross-correlation operation and the Least Median of Squares (LMedS) approach [26]. Then the world coordinate system is established based on projective reconstruction technique and the four coplanar points specified by user in the two control images respectively.

Once the initialization stage is completed, the online registration can start. Feature correspondences detected from the reference images are tracked in the current frame benefiting from the tensor of previous frame and the normalized cross-correlation operation. The trifocal tensor is calculated robustly with the feature triplets based on the method discussed in Section 6. Then the four coplanar points specified by user are transferred into the current image using the calculated tensor, and the homography between the current frame and the world plane is recovered via the correspondence of these four points. Finally the registration matrix is calculated using the above homography and the virtual objects are rendered on the real scenes using OpenGL.

In the above two stages, we assume that the intrinsic and distortion parameters of the camera are known and do not change.

4 ESTABLISHING WORLD COORDINATE SYSTEM

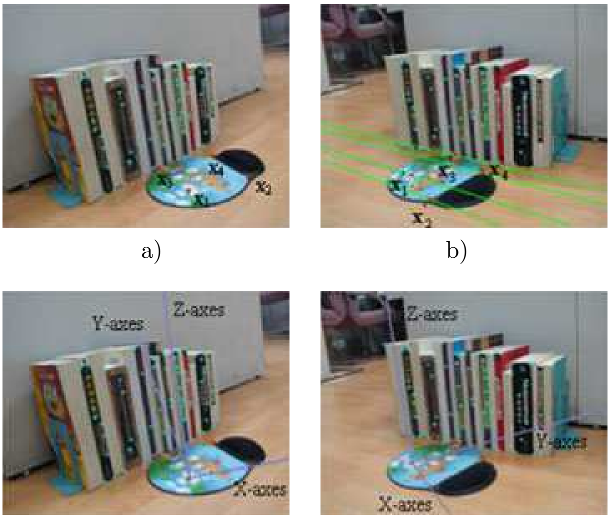
Before establishing the world coordinate system, we assume that the epipolar geometry between the two reference images is known and the projective matrices P and P' of the two reference views have also been calculated through Equation (5). Then, based on the principle of projective reconstruction, we can compute a point's 3D projective coordinates X using its projections on the two reference images as discussed in Section 2.2.

The next step is to specify four coplanar points $x_i = (x_i, y_i, 1)^T$ $(i = 1, \ldots, 4)$ in each of the two reference images, respectively, to establish the world coordinate system. Previous work [7, 8, 9] has the restriction that these four points should form an approximate square, and the origin of the world coordinate system is set to the center of the square defined by the four known points, the X-axes and Y-axes are the direction of two different parallel sides, respectively. However, this method still needs some special square planar structures to help specifying the planar points accurately in the control images.

From the discussion in Section 2.1 we know that the registration matrix can be computed from the homography between current frame and the world plane, and to get this homography, we need only four coplanar but noncollinear points of the world plane and their projections on the current image. Motivated by this property, we propose a new method to define world coordinate system without the need of special square structures, and the only limitation of our approach is that the four specified coplanar points should not be collinear. Let $X_i = (X_i, Y_i, Z_i, 1)^T$ $(i = 1, \ldots, 4)$ be the projective 3D coordinates of the four specified points, the origin of the world coordinate system will be the X_1, the X-axes will be the direction of the vector $\overrightarrow{X_1X_2}$, the Z-axes will be the vertical direction of the plane defined by the above four 3D points, and the Y-axes will be the cross product of Z-axes and X-axes. To improve accuracy, when one point has been specified in a reference image, its epipolar line in another image is drawn to limit the searching area of the correspondence in this image, because the correspondence is limited on this epipolar line according to the property of epipolar geometry. Figure 2 gives an illustration of our method.

5 FEATURE TRACKING

For each incoming frame, we must identify which features correspond to which in the reference images. A candidate method is to track these features frame by frame using narrow baseline feature tracking techniques like Lucas-Kanade tracker [27]



a) b)

c) d)

Fig. 2. Setting up the world coordinate system. a) and b) are the two reference images in which four noncollinear points are specified, respectively, to set up the world coordinate system. To improve accuracy, epipolar lines are drawn on the second image to limit the searching area of the point corresponding to the specified points in the first image. c) and d) are the images with the established world coordinate system.

and so on. However, these methods suffered from losing features, and are prone to introducing some wrong matches. This is especially true in the case of features going out of the field of view or occluded by users and some scene objects. Thus the valid matches will become less and less during the tracking process, which will finally result in failure of registration.

To overcome the above problems, we propose a normalized cross-correlation based method to match the features between current and reference images directly. We assume that the tensor of previous frame has been calculated accurately. With this tensor, we transfer the corresponding feature points detected from the two reference images onto the live image. The correspondence is identified by searching in a small area surrounding the transferred point for a point that correlates well with one of the two reference matched points. Figure 3 gives a perspicuous illustration of the proposed method. By the above method, not only do we fulfil the task of establishing feature correspondences, but also constitute a NCC based criterion to evaluate the quality of point matches. In deed, this is a very important precondition of the modified RANSAC algorithm we used to estimate trifocal tensor in Section 6.

However, the proposed method has the limitation that the initial camera position should be close to one of the two reference images so that we do not have to solve wide baseline matching problem for the first frame in case of the absence of initial tensor.

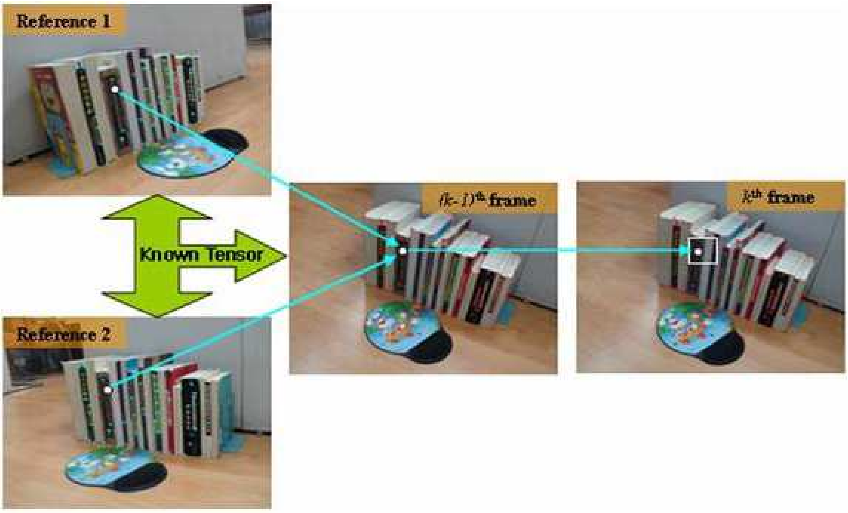


Fig. 3. Tracking features. A searching window is fixed in the current frame using the previous tensor. The initial correspondence is established by searching in this window for a point that has the maximal normalized cross-correlation score with one of the two references matched points.

6 ESTIMATING TRIFOCAL TENSOR USING MODIFIED RANSAC

We now turn to the problem of calculating trifocal tensor using feature triples obtained from the previous section. The trifocal tensor has 27 elements, 26 equations are needed to calculate a tensor up to scale. Each triplet of point matches can give four independent linear equations for the entries of the tensor; accordingly, 7 points are needed to compute the trifocal tensor linearly. The above method is called linear algorithm or normalized linear algorithm when input data is pretreated. The drawback of these algorithms is that they do not take into account the internal constraints of the trifocal tensor and cannot necessarily yield a geometrically valid tensor. Algebraic minimization algorithm may be a good choice to obtain a valid tensor. However, due to using all available triples, this method is prone to being affected by the presence of mismatches (outliers). To overcome the disturbance of the mismatches, some scholars have brought forward the 6-point RANSAC method. This method has the capability of generating a precise tensor even in the presence of a significant number of outliers. The shortcoming of RANSAC is that its computational complexity increases dramatically with the number of correspondences and proportion of mismatches. Moreover, 6-point RANSAC does not always work well in a multi-planar scene [28]. If the selected 6 points are accidently coplanar, the resulting tensor can only provide correct geometry for one plane. Consequently, to achieve the expected performances in both time and precision, we propose a modified RANSAC based method to calculate trifocal tensors which takes full advantage of normalized linear algorithm and algebraic minimization algorithm.

6.1 Sampling with Modified RANSAC Algorithm

Standard RANSAC algorithm treats all matches coequally and extracts random samples uniformly from the full set. It can be viewed as a black box that generates N tentative correspondences, i.e. the error-prone matches established by comparing local descriptors. In our experiment, we find that the correspondences with higher normalized cross-correlation score are more likely to be inliers than the lower ones. Motivated by this property, we propose a modified RANSAC algorithm in which samples are semi-randomly drawn from a subset of the matches with the highest cross-correlation score, and the size of the hypothesis generation set is gradually increased. In our method, the size of the set of potential matches has very small influence on its speed, since the solution is typically found early, when samples are taken from a smaller set. In fact, our method is designed to draw the same samples as standard RANSAC algorithm, but only in a different order. The matches more likely to be inliers are tested prior to the others; thereby, the algorithm can arrive at termination criterion and stop sampling earlier. The experiments presented in Section 7.2 demonstrate that this method is valid and can reduce sample times to a large degree compared with the standard RANSAC.

In our method, the set of K potential triples is denoted as N_K. The data points in N_K are sorted in descending order with respect to the normalized cross-correlation score s.

$$n_i, n_j \in N_K : i < j \Rightarrow s(n_i) \geq s(n_j) \tag{12}$$

A set of k data points with the highest score is represented as N_K. Then, the initial subset contains the 7 top-ranked matches that can give 26 equations needed in normalized linear algorithm. If all of the valid samples from the current subset $N_m = (n_1, n_2, \ldots, n_m)$ have been tested, then the next subset is $N_{m+1} = (n_a, n_2, \ldots, n_m, n_{m+1})$, and the following samples consist of n_{m+1} and the 6 data points drawn from N_m at random.

6.2 Algebraic Minimization Algorithm

The standard algebraic minimization algorithm takes the following steps [24]:

1. From the set of feature triples, find an initial estimate of the trifocal tensor using normalized linear algorithm by solving a set of equations of the form $At = 0$, where A comes from Equation (11), t is the vector of entries of tensor T_i^{jk}.
2. Find the two epipoles $e'(a_4)$ and $e''(b_4)$from the initial tensor as the common perpendicular to the left null-vectors of the three T_i.
3. According to Equation (10), construct the 27×18 matrix E such that $t = Ea$, where a is the vector representing entries of a_i^j and b_i^k.

4. Compute the tensor by minimizing the algebraic error $\|AEa\|$ subject to $\|Ea\| = 1$.
5. Find an optimal solution by iteration over the two epipoles using an iterative method like Levenberg-Marquardt algorithm.

To get a fast non-iterative algorithm, we omit the last iteration step in our algorithm. The experiments presented in Section 7.1 prove that the negative influence of this predigestion is very slight and can be ignored.

6.3 Estimating Method

The following steps give the outline of our modified RANSAC based algebraic minimization algorithm.

1. From the sample given by modified RANSAC algorithm, compute a candidate tensor using the normalized linear algorithm.
2. Reproject all the matches of the two references on to the current frame using the candidate tensor and Equation (11).
3. If the number of inliers is smaller than the predefined threshold T (varying with different environment), then generate a new sample using modified RANSAC and repeat the above steps.
4. If the number of inliers is greater than T, then re-estimate the tensor using these inliers and the method described in Section 6.2 and terminate.

In our method, the criterion to judge an outlier is that the distance between the reprojection and the detected point in current frame is greater than 3 pixels.

7 OPTIMIZING REGISTRATION MATRIX

After getting the needed tensor, we can calculate the registration matrix using the method discussed in Section 2. However, we find that the system suffered from the problem of jitter. This problem is especially acute when current position is far away from the reference images. To overcome the above problem, we use nonlinear least square method (Levenberg-Marquardt algorithm) and Tukey M-estimator [29] to optimize the initial registration matrix to get a more stable result.

$$\min_{R,T} \sum_{i}^{n} \rho(r_i) \tag{13}$$

where ρ is the Tukey M-estimator and $r_i = \|x_i - \lambda K[R|T]X_i$ is the re-projection error. For Tukey M-estimator $\rho_{Tuk}(x)$ is computed as:

$$\rho_{Tuk}(x) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{x}{c}\right)^2\right)^3\right] & \text{if } |x| \le c \\ \frac{c^2}{6} & \text{if } |x| > c \end{cases} \tag{14}$$

where c is a threshold chosen with respect to the standard deviation of the data.

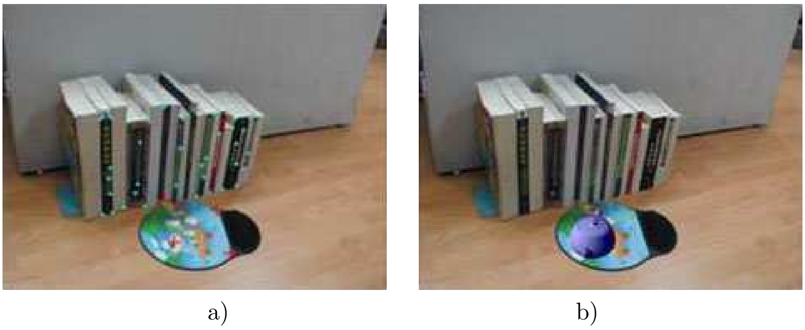
By minimizing the residual sum, the camera rotation and translation matrix R and T for the current frame can be estimated accurately. In our algorithm, M-estimator is initialized with the camera pose of the previous frame.

8 EXPERIMENTAL RESULTS

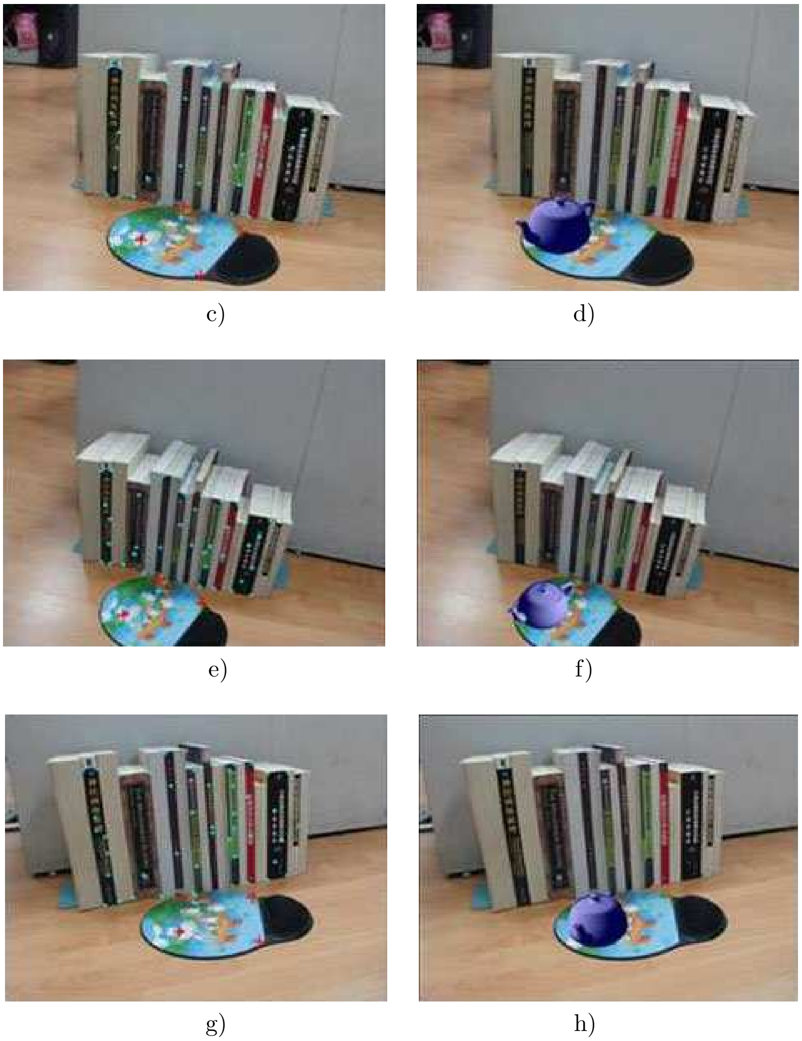
The proposed method has been implemented in C using OpenGL and OpenCV [30]. The video sequence is captured with a CCD camera. To calibrate this camera, we make use of the camera calibration toolbox of OpenCV. We put a checkerboard in front of the camera and capture several sample images. The recognized corners in the checkerboard images and the known geometry of the checkerboard allow us to obtain the parameters required in our system. Indubitably, the speed of our system varies according to different conditions appearing in the video sequences and the complexity of virtual objects. On the average, our system can run at 10 fps with 320 × 240 pixel images on a DELL 530 Workstation (OS: Windows 2000, CPU: 1.8 GHz × 2).

Two experiments are carried out to demonstrate the validity of the proposed approach.

In the first experiment, two control images are first selected with the camera placed at different positions. An initial set of feature matches is made by normalized correlation, and a classical robust approach, called the Least Median of Squares, based on the technique described in [26], is used to estimate the epipolar geometry between these two reference images. Then, more matches can be found by stereo matching with the computed epipolar geometry. In our work, 73 point matches are obtained finally after repeating the above steps 4 times. Projective camera matrices for the two views are computed next. Four pairs of points are then specified in the two reference images to establish the world coordinate system. Using the tensor obtained with the method discussed in Section 6.3, the four specified points are transferred during the entire tracking process. Figure 4 shows some results of the augmented sequence.



a) b)



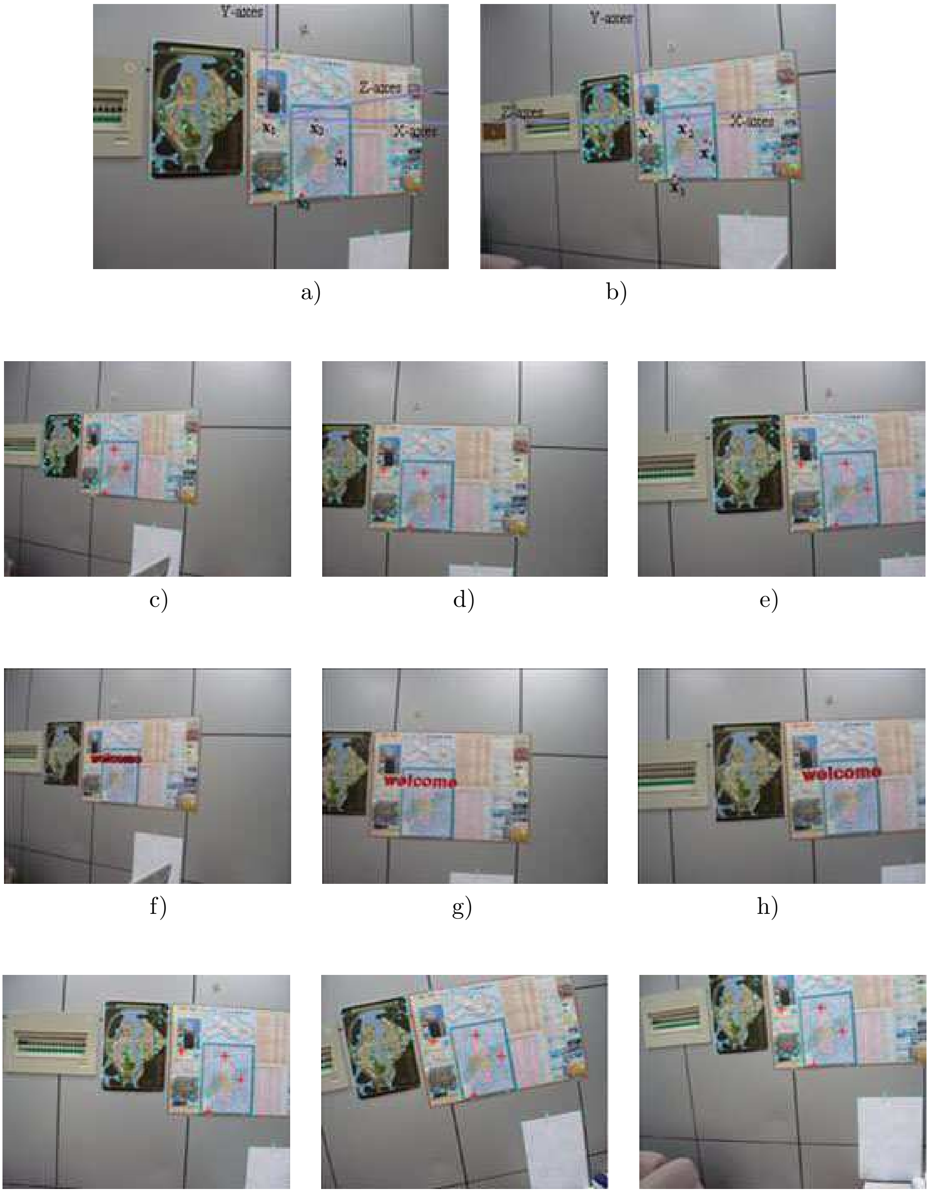
c) d)

e) f)

g) h)

Fig. 4. Examples in experiment 1. a), c), e) and g) are the 45th, 86th, 153rd and 257th frames of the input video sequence with the inliers used to calculate the needed tensor and the reprojection of the specified point marked with the symbol "+", respectively. b), d), f) and h) are the corresponding registration images.

In the second experiment, 131 matches are extracted from the two reference views. A 3D virtual word "welcome" is successfully augmented on the wall over thousands of consecutive frames. Figure 5 shows some images of the augmented sequence.

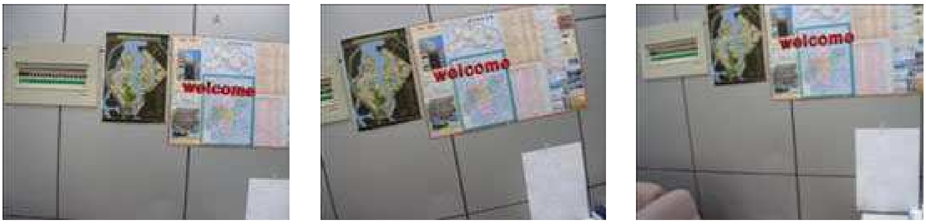


a) b)

c) d) e)

f) g) h)

i) j) k)



l) m) n)

Fig. 5. Examples in experiment 2. a) and b) are the two reference images with the feature matches and the established world coordinate system. c), d), e), i), j) and k) are the 23rd, 76th, 133rd, 150th, 251st, 364th frames of the input video sequence with the inliers and the reprojection of the specified point marked with the symbol "+", respectively. f), g), h), l), m) and n) are the corresponding registration images with a virtual 3D word "welcome".

8.1 Tracking Accuracy

In order to evaluate the registration accuracy in our method, the results of the above two experiments are recorded. The validation of our results is obtained by a comparison with one of the commercially available 3D magnetic (miniBIRD). This highly accurate system can output both position and orientation and has an accuracy of 0.5 in orientation and 1.8 mm in position. The wired sensor is installed on the CCD camera. Poses from both systems are simultaneously acquired. Tracking results (Translation of Y-axes) are shown in Figures 6 and 7. Position errors in Y-axis are all less than 10 mm compared with miniBIRD in both experiments. These results demonstrate the accuracy of the proposed method.

8.2 Sample Times

Using the video sequence obtained from the second experiment, we also compare the sample times between the modified RANSAC and the standard RANSAC algorithm. With the method discussed in Section 5, we typically generate sets of initial matches which are already 80 % correct. The threshold T for these two algorithms is set to 65 % of the matches in initial sets. The comparision of the sample times is given in Figure 8. The average and maximum sample times of the standard RANSAC in the first 500 frames are 10.3 and 79 respectively, which is 4.29 and 5.64 times as in our method. The processing time of the standard RANSAC algorithm in this experiment varies from 0.016 s to 0.172 s, which leads to the frame rate drifts between 4 and 12 fps; on the other hand, the processing time of our method is always within 0.047 s and the frame rate is about $9 \sim 12$ fps. This experiment proves that the modified RANSAC algorithm used in our method is more stabile and can reduce the computation complexity significantly, which makes our system suitable for online implementation.

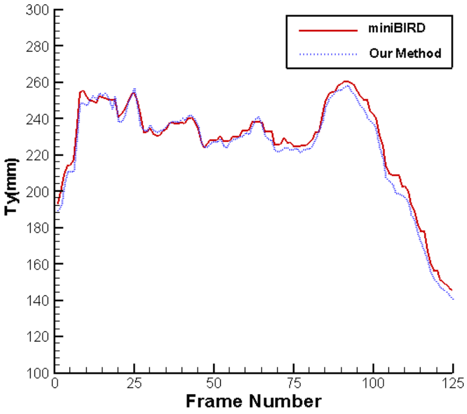


Fig. 6

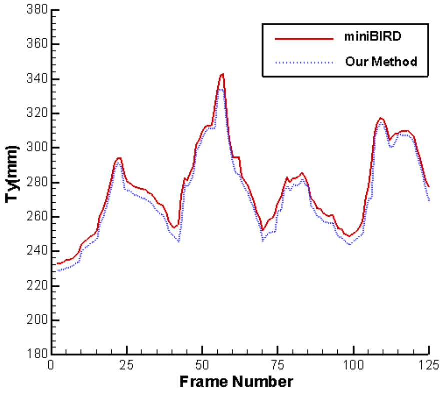


Fig. 7

9 CONCLUSIONS

In this paper, we presented a markerless registration method for augmented reality systems based on online estimation of trifocal tensor using natural features. To establish the world coordinate system, we relax the restriction that the four specified points must form a square. This method casts off the requirements of special square planar structures and really enhances the usability of our system. To calculate trifocal tensor precisely, we put forward a modified RANSAC based algebraic minimization algorithm. While improving the accuracy, this method also reduces the computation complexity to a large degree, which really ameliorates the efficiency of our system.

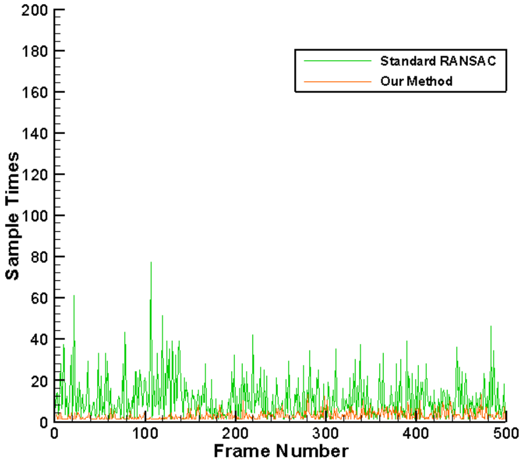


Fig. 8

However, there are still some limitations. In the current system, virtual objects are usually superimposed on the video image without using depth information from the real scenes, so real objects are always occluded by virtual objects. This problem reduces the impression that the virtual objects are part of the real world scene and affects the user's recognition of the geometrical relationship between real and virtual objects. According to the property of trifocal tensor, we can transfer foreground pixels of the reference images to the current frame to realize the reprojection of the real objects. If the depth information of these foreground pixels can be simultaneously calculated in real-time, then we can achieve occlusion between real and virtual objects easily. This is the direction of our future work.

REFERENCES

[1] Soler, L.—Nicolau,S.—Schmid, J.—Koehl, C.—Marescaux, J.—Pennec, X.: Virtual Reality and Augmented Reality in Digestive Surgery. In Proc. of ISMAR, 2004, pp. 278–279.

[2] Brown, D. G.—Coyne,J. T.—Stripling, R.: Augmented Reality for Urban Skills Training. IEEE Virtual Reality, 2006, p. 35.

[3] White, M.—Mourkoussis, N.—Darcy, J.—Petridis, P.: ARCO – An Architecture for Digitization, Management and Presentation of Virtual Exhibitions. In Proc. of ICG, 2004, pp. 622–625.

[4] Uematsu, Y.—Saito, H.: AR Registration by Merging Multiple Planar Markers at Arbitrary Positions and Poses via Projective Space. In Proc. of 15th International Conference on Artificial Reality and Telexistence, 2005, pp. 48–55.

[5] Mark, A. L.—Andrei, S.: Magnetic Tracker Calibration for Improved Augmented Reality Registration. Presence: Teleoperators and Virtual Environments. MIT Press, Vol. 6, 1997, No. 4, pp. 532–546.

[6] Ogris, G.—Stiefmeier, T.—Junker, H.: Using Ultrasonic Hand Tracking to Augment Motion Analysis Based Recognition of Manipulative Gestures. In Proc. of ISWC, 2005, pp. 152–159.

[7] Pang, Y.—Yuan, M. L.—Nee, A. Y. C.—Ong, S. K.: A Markerless Registration Method for Augmented Reality based on Affine Properties. In Proc. of AUIC, 2006, pp. 25–32.

[8] Yuan, M. L.—Ong, S. K.—Nee, A. Y. C.: Registration Based on Projective Reconstruction Technique for Augmented Reality Systems. IEEE Trans. on Visualization and Computer Graphics, Vol. 11, 2005, No. 2, pp. 254–264.

[9] Yuan, M. L.—Ong, S. K.—Nee, A. Y. C.: Registration Using Natural Features for Augmented Reality Systems. IEEE Trans. on Visualization and Computer Graphics, Vol. 12, 2006, No. 3, pp. 569–580.

[10] Simon, G.—Berger, M.: Reconstructing While Registering: A Novel Approach for Markerless Augmented Reality. In Proc. of ISMAR, 2002, pp. 285–294.

[11] Simon, G.—Berger, M. O.: Real Time Registration Known or Recovered Multi-Planar Structures: Application to AR. In Proc. of BMVC, 2002, pp. 567–576.

[12] Uematsu, Y.—Saito, H.: Vision Based Registration for Augmented Reality Using Multi-Planes in Arbitrary Position and Pose by Moving Uncalibrated Camera. In Proc. of Mirage 2005, pp. 111–119.

[13] Bartoli, A.—Tunzelmann, E.—Zisserman, A.: Augmenting Images of Non-Rigid Scenes Using Point and Curve Correspondences. In Proc. of CVPR, Vol. 1, 2004, No. 1, pp. 699–706.

[14] Comport, A. I.—Marchand, E.—Chaumette, F.: A Real Time Tracker for Markerless Augmented Reality. In Proc. of ISMAR, 2003, pp. 36–45.

[15] Vacchetti, L.—Lepetit, V.—Fua, P.: Stable Real-Time 3D Tracking Using Online and Offline Information. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, 2004, No. 9, pp. 1391–1391.

[16] Vacchetti, L.—Lepetit, V.—Fua, P.:. Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking. In Proc. of ISMAR, 2004, pp. 48–56.

[17] Lepetit, V.—Lagger, P.—Fua, P.: Randomized Trees for Real-Time Keypoint Recognition. In Proc. of CVPR, Vol. 2, 2005, pp. 775–781.

[18] Nister,. D.: An Efficient Solution to the Five-point Relative Pose Problem. IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 26, 2004, No. 9, pp. 756–770.

[19] Nister, D.—Naroditsky, O.—Bergen, J.: Visual Odometry. In Proc. of CVPR, 2004, pp. 652–659.

[20] Davison, A. J.: Real-Time Simultaneous Localisation and Mapping With a Single Camera. In Proc. International Conference on Computer Vision, 2003.

[21] Molton, N. D.—Davison, A. J.—Reid, I. D.: Locally Planar Patch Features for Real-Time Structure from Motion. In Proc. British Machine Vision Conference 2004.

[22] Davison, A. J.—Reid, I. D.—Molton, N. D.—Stasse, O.: MonoSLAM: Real-Time Single Camera SLAM. IEEE Transaction on Pattern Analysis and Machine Intelligence. Accepted for publication, 2007.

[23] Zhang, Z.: Determining the Epipolar Geometry and Its Uncertainty: A Review. Int'l J. Computer Vision, Vol. 27, 1998, No. 1, pp. 161–198.

[24] Hartley, R.—Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, 2000.

[25] Harris, C.—Stephens, M.: A Combined Corner and Edge Detector. In Proc. of Alvey Vision, 1988, pp. 189–192.

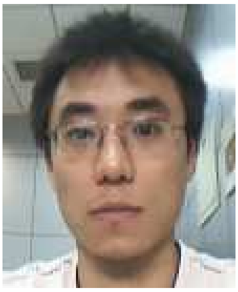
[26] Zhang, Z.—Deriche, R.—Faugeras, O.—Luong, Q. T.: A Robust Technique for Matching Two Uncalibrated Images through the Recovery of the Unknown Epipolar Geometry. Int'l J. Artificial Intelligence, 1995, pp. 87–119.

[27] Lucas, B.—Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In Proc. of IJCAI, 1981, pp. 674–679.

[28] Fraundorfer, F.: Robust Estimation of the Trifocal Tensor and it's Application to Image Matching. Master Thesis. Graz University of Technology, 2001.

[29] Vacchetti, L.—Lepetit, V.—Fua, P.: Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking. In Proc. of ISMAR, 2004, pp. 48–56.

[30] OpenCV: `http://www.intel.com/research/mrl/research/opencv`.



Tao Guan received the Ph. D. degree in Information and communication science from HuaZhong University of Science & Technology in 2008. He is currently doing his postdoctoral researches in Digital Engineering and simulation Research Center, HuaZhong University of Science & Technology. His research interests include computer graphics, computer vision, and augmented reality.

Lijun Li received the B. Sc. and M. Sc. degree from the Shanghai Jiao Tong University in 1997 and HuaZhong University of Science & Technology in 2003 respectively. He is currently a junior researcher in Digital Engineering and simulation Research Center, HuaZhong University of Science & Technology. His research interests include computer graphics, virtual reality, and augmented reality.

Cheng Wang received his Ph. D. degree from Massachusetts Institute of Technology in 1983. He has been a professor of digital engineering in the Digital Engineering and simulation Research Center, HuaZhong University of Science & Technology since 1999. His research interest is in large-scale structure analysis and simulation, virtual and augmented reality. He has published two books and more than 50 papers in refereed journals and conference presentations.