IRSTI 28.17.33 https://doi.org/10.26577/1i32jpcsit2301

Jia Na

Jiangnan University, Wuxi, Jiangsu, China
e-mail: 516306088@qq.com

RESEARCH ON AUGMENTED REALITY TECHNOLOGY

Abstract. With the coming of the 21st century, man is entering into a brand-new multimedia age. Digital Video processing, as the most important, expressive and complex one in multimedia, has progressed quite a lot. An augmented reality system generates a composite view for the user. It is a combination of the real scene viewed by the user and a virtual reality scene generated by the computer that augments the scene with additional information. It is a new technology, which can combine the real world information with the virtual world information seamless.

In this paper, a simple registration method using natural features based on the projective reconstruction technique is proposed. This method consists of two steps: embedding and rendering. Embedding involves specifying four points to build the world coordinate system on which a virtual object will be superimposed. In rendering, the Kanade-Lucas-Tomasi (KLT) feature tracker is used to track the natural feature correspondences in the live video. The natural features that have been tracked are used to estimate the corresponding projective matrix in the image sequence.

Key words: augmented reality, registration, projective reconstruction, natural feature tracking.

1 Introduction

In recent years, the research of augmented reality technology, which combines virtual reality and real scenes, has attracted much attention. There are many international conferences dedicated to this topic, such as the IEEE Workshop on Augmented Reality (AR) and the International Symposium on Mixed Reality, etc. Some well-known laboratories are also working on this project, including the Japanese Mixed Reality SystemsLab and the German Arvika Consortium.

The original virtual reality technology uses an opaque helmet or goggles (HMD, Head Mounted Device) to present a fully computer-synthesised virtual world to the viewer. In this technology, the lack of realism of the computer-synthesised scenes and the requirements for the computer peripherals limit its application to some extent [1]. The Robotics Institute, led by Professor Takeo Kanade of Carnegie Mellon University in the United States, and the research group of Professor Yuichi Ohta of Tsukuba University in Japan have developed a system called virtualized reality. In previous virtual reality technology, the entire three-dimensional scene was synthesised by the computer, and even the three-dimensional objects in the scene consisted solely of models simulated by the computer. Virtual reality (VR) consists of the computer recording and processing events that take place in the real world so that it is closer to reality. The system installs fixed cameras with different viewing angles on the football field, uses a computer to process multiple video streams with different viewing angles that are recorded simultaneously, and performs a three-dimensional reconstruction of the real scene. In this way, observers can watch the game from any angle that differs from the cameras. They can pretend to fly over the pitch. Alternatively, they can stand in the middle of the football field or watch the game from the position of a specific player [2]. The 3D Image Overlay System, also developed by Carnegie Mellon University, combines computer-generated images with three-dimensional real scenes. People can not only observe the real scene without hindrance, but also get more useful information from the images overlaid by the computer.

This augmented reality technology can be used to improve the surgical environment and remote diagnosis. Similar to virtual reality technology, augmented reality technology also requires the help of HMD devices. The difference is that these

© 2023 Al-Farabi Kazakh National University

devices do not rely on virtual reality to provide more effects that are useful.

After learning about the development stages of virtual reality, virtualized reality, and augmented reality, the famous scientist Professor Azuma and others recently summarized these technologies as mixed reality in the journal IEEE Computer Graphics and Applications. Through these technologies, the three-dimensional "reality" appears on the glasses. Basic research on this technology is ongoing, and applications are expanding. However, a large number of existing video images (e.g., movies, TV and DVD, etc.) are two-dimensional image sequences of real scenes. With the development of visual geometry and camera self-calibration technology, it is hoped that the desired 3D virtual objects can be inserted into the existing video sequence of the real scene captured by the camera. In order to insert the virtual object, sufficient three-dimensional information must be obtained from the existing two-dimensional video sequence, and the internal and external parameters of the camera are critical. While in the past it was assumed that this was possible, it was very difficult to accurately recover these parameters from video sequences obtained with arbitrary camera motion, i.e., accurate self-scaling techniques. In recent years, with the rapid development of visual geometry, camera self-calibration technology, and 3D reconstruction from video sequences, the research conditions for inserting 3D virtual objects into real video sequences have gradually matured, creating the field of mixed computer reality (mixed reality). This hybrid video technology, in which virtual three-dimensional objects are inserted into real video sequences, is called augmented video). Compared with other technologies, the research of inserting virtual 3D objects into existing video sequences has much more attractive advantages and offers a wide range of possible applications. It is free from the constraints associated with using HMD devices and installing a large number of cameras in real scenes. The video sequence obtained with this technique can be captured with a single camera, and its position and motion are not limited, including the video sequence of real scenes captured with a handheld camera. The captured video sequence is only used to reconstruct the three-dimensional coordinate information of the scene, determine the internal and external parameters of the camera, and insert virtual three-dimensional objects to create a new video sequence with mixed scenes. This is undoubtedly the advantage of mixed reality technology. A groundbreaking achievement and development.

Based on the research in visual geometry and 3D reconstruction, the research group of Professor Andrew Zisserman from the Computer Department of Oxford University has done some research to improve video technology in recent years and made some progress [3]. They conducted an experiment in which a 3D object was inserted into a known video sequence. In the newly generated enhanced video sequence with 40 frames, the virtual object and the 3D scene in the known video sequence are merged, which gives a better visual sense of coexistence, but the question of how to reduce the jitter of the inserted object compared to the original scene is still unsolved and needs further research. The research results show that there is a discrepancy of several pixels between the inserted 3D virtual object and the 3D scene reconstructed from the known video sequence. The problem of occlusion, which affects the visual impact of the final enhanced video sequence, was not addressed in the study. When projecting the added 3D virtual object into the existing video sequence, it is necessary to determine whether the virtual object is occluded by a certain part of the real scene, i.e., the relative distance between each object in the scene and the camera must be calculated. To do this, the relative distance between each object in the scene and the camera must be calculated, and then the edge of the occluded area must be determined so that a new, enhanced video sequence can be created correctly.

AR is an important branch of virtual reality technology VR and also an emerging research focus. Virtual reality is defined as a three-dimensional interactive virtual scene generated by a computer that gives the user the feeling of being immersed in it. However, real-world scenarios are quite complex and contain a large amount of data. With available technology, it is difficult to fully describe them, regardless of the computer's capabilities and the accuracy of the graphics generated. In many cases, people need a better understanding of the real scene, so it is not necessary to generate all the scenes with the computer. In this context, augmented reality technology is used. Augmented reality consists of the real scene that the user sees and the computer-generated virtual scene superimposed on the real scene. The virtual scene enhances the real scene and improves people's perception and understanding of the real scene. The ultimate goal of augmented reality is to create a scene that is completely integrated with the real scene and the virtual scene, so that the user does not feel what is real and what is virtual, but thinks that what they are seeing is a completely real scene.

Augmented reality technology is characterized by the combination of virtual and real reality, real-time interaction, and three-dimensional registration. Augmented reality technology can use the existing real world to provide a composite visual effect to the user: The real world observed by the user is merged with the virtual scene generated by the computer. As the real scene moves, the virtual objects also sound and change, as if these virtual objects really exist in the real scene. Ideally, the virtual objects can also interact and interact with the user and the real objects in a natural way. The difficulty in developing an augmented display system is determining the exact position and attitude of the camera in relation to the real world in real time, so that the virtual scene can be seamlessly integrated with the real world (3D registration). The core of augmented reality is the integration of the real and virtual worlds. In the existing technology, optical technology and video technology can be used to realize the fusion of real and virtual [4]. The basic principle of optical technology to achieve the fusion is to superimpose the two by semi-transparent and semi-reflective lenses. The principle of video technology to achieve the fusion is to superimpose the image of the virtual world with the real image in the graphics processor by the camera, and then display it uniformly on the screen. The screen can be a desktop graphics monitor or a transparent helmet-mounted display (HMD). If the method of implementing augmented reality is different, the hardware composition of the augmented reality system is also different.

This article presents four planar specified points form a region where the virtual objects will be superimposed, and attempts to convert these points using natural features, and the specified region can also be random.

2. KLT Feature Tracking System

Tracking of natural features is a topical research in computer vision [5-9], and this section offers a sketch of the classical KLT feature tracking system [6].

When the camera moves ,the change in frame intensity of an image can be seen as a function with three variables, *I(x,y,t)*. The change in frame intensity can then represent a sequence of images. *i(x)* This function is sometimes abbreviated to *I(t)* in later sections. The function *I(x,y,t)* satisfies formula (1).

$$I(x+dx, y+dy, t+\tau) = I(x, y, t). \qquad (1)$$

The displacement of a point in $I(x, y, t)$ at times t and $t + \tau$ is defined as$d = (dx, dy)$. A critical problem in determining the displacement $d = (dx, dy)$ of a point from one frame to the next is that a single pixel point cannot be used as a tracking point; the value of this pixel point may change due to noise and it is difficult to identify adjacent pixels.

The image gained at $t + \tau$ time is defined as the image *J* and equation (1) can be redefined as

$$J(x+d) = I(x+dx, y+dy, t+\tau).$$

It is indicates that the point x only translates moves from the first image *I* to the second image *J*. For a given viewport w, the problem of determining its motion para-meters is to minimize its difference points:

$$\varepsilon = \iint [J(x+d) - I(x))]^2\, w(x)\, dx, \qquad (2)$$

Of these, *w(x)* is the weight function and is usually set to 1.

Equation (2) can be reorganized by using the Taylor series as follows.

$$Zd = e, \qquad (3)$$

Of these, Z is a 2 × 2 matrix $z = = \int \int_m g(x) g^t(x) w(x) dx$ and $g(x) = \begin{bmatrix} \frac{\partial}{\partial x}(I+J) \\ \frac{\partial}{\partial y}(I+J) \end{bmatrix}$, e is a 2 × 1 vector

$$e = 2 \int \int_m [I(x) - J(x)] g(x)(x) w(x) dx. \qquad (4)$$

Using the intensity slope, Z and e can be written in the following form respectively

$$Z = \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y \end{bmatrix};\ e = -2\sum_w Id \begin{bmatrix} g_x \\ g_y \end{bmatrix};$$

Where the g_x and g_y are the partial derivative derived from $I(x,y,t)$, Id is the differrence in intensity between images *I* and *J*.

To minimize the difference between the two viewports, the displacement *d=(dx,dy)* can be reckoned by using Newton's iterative method. Prioritize selecting good features before performing the KLT feature tracker. In fact, the symmetric matrix in (3) is the covariance matrix derived from the image, which states that the image structure is distributed over a small area [5, 6, 10]. A small eigenvalue of *Z* corresponds to a relatively constant intensity in the region. The distribution of eigenvalues of Z predicts the calculated value of the optical flux at a point, so it is good for selecting of feature points. In practice, when one of the small eigenvalues of Z is large enough to be larger than a given lower bound λ_0, then the points can be accepted as features when:

$$min(\lambda_1,\lambda_2) > \lambda_0. \qquad (5)$$

In the above discussion, only a purely translational movement was considered. When large viewpoint changes and tracking of long sequences are involved, the affine motion model can be used to improve the KLT feature tracker. The affine motion model [5, 11, 9, 10] is as follows:

$$J(Ax + d) = I(x). \qquad (6)$$

Dissimilarity can be redefined as follows:

$$\varepsilon = \iint [J(x+d) - I(x))]^2\, w(x)\, dx,. \qquad (7)$$

For distorted linear images, the affine transformation takes more time, which affects the real-time performance of the system AR. In fact, it is sufficient to use a purely translational motion model in the registration algorithm described in this paper, since only at least 6 natural points need to be tracked to estimate the corresponding projection matrix in the live video, and then use this matrix to track those to be virtually overlaid The particular area of the object.

3. Fundamentals of registration

Image points and 3D points are represented by the flush vectors $m = (u,v,1)^r$ and $M = (X,Y,Z,1)^r$respectively. The connection between a three-dimensional point M and its image projection m is generally as follows:

$$\rho m = A[R|t]M, \qquad (8)$$

Where p is an random coefficient, R and t show the rotation matrix and displacement vector of the camera comparing the world coordinate system respectively, and are generally referred to as exterior parameters, A is called to as the camera intrinsic matrix.

In any two camera systems, there exists the pair-polar geometry [12]. The fundamental matrix F generalizes this pair-polar geometry, and F is a 3x3 matrix of rank 2. It has infinite projection bases in order to satisfy the pair-polar geometry. F acts as the product of the antisymmetric matrix $[e' \ _x$, and the matrix$M$, F=$[e' \ _x M$, then two projection camera matrices can be chosen [12, 13]:

$$P = [I|0], P' = [M|e'], \qquad (9)$$

P And P' define a unique projection space, and given a pair of matching points (m_{1i}, m_{2i}) between two images, their corresponding 3D projection coordinates M_i , can be derived from the equations $sm_{1i} = PM$ and $s'm_{2i} = P'M^\tau$ by the mean of least squares or least eigenvalue, and the s and s' are any two scalars. For them_{ki}th image, the relationship between the image coordinates M_i and its 3D projection coordinates is as follows：

$$\rho m_{ki} = P^k M_i, \qquad (10)$$

$P^k$is a corresponding projection matrix of the k th image with 11 unknown parameters . If there exists at least 6 pairs of points （m_{ki}, M_k), then the projection matrix P^kcan be estimated. On the other hand, the projection matrix can reflect each 3D projection coordinate to a 2D projection point. So, if a 3D projection point is known in the initial stage, its projection can be calculated during the tracking process by using the estimated matrix. Given by equation (10) for n pairs of points (m_{ki}, M_k), the following linear equation must be satisfied：

$$Ap = 0 \tag{11}$$

where A is a 2n × 12 matrix and p represents a vector of all parameters of P^k, which can be concluded by the least eigenvalue way.

4. Registering Algorithm

This section introduces a registration algorithm based on the KLT feature tracker and projection reconstruction techniques. The registration involves two steps: embedding and reproduction. In fact, the embedding phase is used to initialize the AR system, and for the sake of convenience, notes the natural features acquired in the Kth image at this series of K moments as $NF(t_k)$, and the natural features in $number\big(NF(t_k)\big)$, are noted as lost $lost\big(NF(t_k)\big)$, is possible to be lost, so that we can found:

$$number\big(NF(t_0)\big) \geq \\ \geq number\big(NF(t_1)\big) \geq \\ \dots \geq number\big(NF(t_k)\big) \tag{12}$$

Two control images, namely$I(t_0)$ and$I(t_1)$, are selected before the embedding. Firstly, the natural features are gained from the first image $I(t_0)$using equation (5). Throughout the upgrading process, we track the natural features in the image $I(t_1)$ with the KLT feature tracker and use them as reference points, and these natural features, are those points in the image $I(t_0)$, that correspond to the natural features $NF(t_0)$that have been detected. Usually, a natural feature corresponds to a distinguishable physical point, but possibly some features do not correspond to physical points (in [12] such features are called "bad" points). For example, in an image of a tree, a horizontal branch in the foreground can cross a vertical branch in the back view, and we may regard the crossover point as a natural feature, however, this crossover point in the image is not the point in the real physical image. The method described in [12, 13] is next used to compute the fundamental matrix on the basis of the points in $NF(t_1)$and their counterparts in $NF(t_1)$. We can have two projection camera matrices P and P* of the control image, by which we can calculate the 3D projection coordinates M.

Next, four plane points $\{x_{0i}\}(i = 1, \dots\dots, 4)$ are specified at the place where the virtual object will be overlaied, the specified four points form the world coordinate system for the x and y axes, and the origin of the coordinate system is at the center of the approximate square formed by the four points, and the z-axis is the vertical direction of the xy plane [14][15]. The four specified plane points can be random, in other words ， they are not necessarily natural features, and when the four matching points are specified, in order to determine the positions of these points in the other images, it is necessary to calculate their associated projected 3D coordinates$\{x_{0i}\}(i = 1, \dots\dots, 4)$, and thus the position and structure of a virtual 3D object can be determined.

After the initial embedding process, the next step is to convert the specified area (four points) on which the virtual object will be superimposed as the user moves. For the kth image in the video sequence, the natural features tracked by the KLT feature tracker and corresponding to$NF(t_k)$6) can be estimated based on the corresponding 3D coordinates $\{\ M_i\ \}$ computed in the embedding phase, using the estimation method in Section 2 to estimate the projection matrix P^k. Since the 3D projection coordinates of the four specified points have been calculated in the embedding stage, we can calculate the projection of these 3D points based on $NF(t_k)$ and their corresponding projected 3D coordinates (i.e., the natural features of the kth map to estimate the corresponding projection matrix.

The main purpose of the algorithm is to transform the specified four points using the natural features that have been tracked based on the projection reconstruction technique, and then R_k and t_k can be easily computed by using the estimation method of ARToolKit.

5. Experiment

The experiment was performed using Visual C++ on Intel Pentium CPU 3.00 GHz, memory 1.00 GB. The image sequence was captured with a IEEE 1394 Firefly camera, and the image size is 640*480. The errors described in these experiments are RMS errors. Although this is an indirect measure of the value of the estimated results, it provides a good quality index of the correspondding projection matrix in the presence of points outside the image. We compare two experiments with normalized data and with raw natural tracking points as input data.

Here, two experiments are conducted indoors. First, two control images are selected where the cameras are at different positions. In the first image, natural features are detected and then tracked using the KLT feature tracker. The natural features in the second control image are the same as those in the first image. The tracked natural features and their corresponding points in the first control image are then used to calculate the fundamental matrix, then the camera projection matrices for the two images are calculated, and finally four matching points in the two control images are determined. To improve accuracy, you can use bipolar geometry to specify four points in the second image that correspond to the four points specified in the first image. Then the three-dimensional projected coordinates of the natural feature and the four specified points are calculated. Throughout the tracking process, the four specified points can be transformed using the registration algorithm. As can be seen in Figure 1, the viewing angle of the camera at the four points marked with the symbol "+" on the tabletop changes greatly from Figure 1(a) to Figure 1(f), which shows that the proposed registration algorithm is effective.

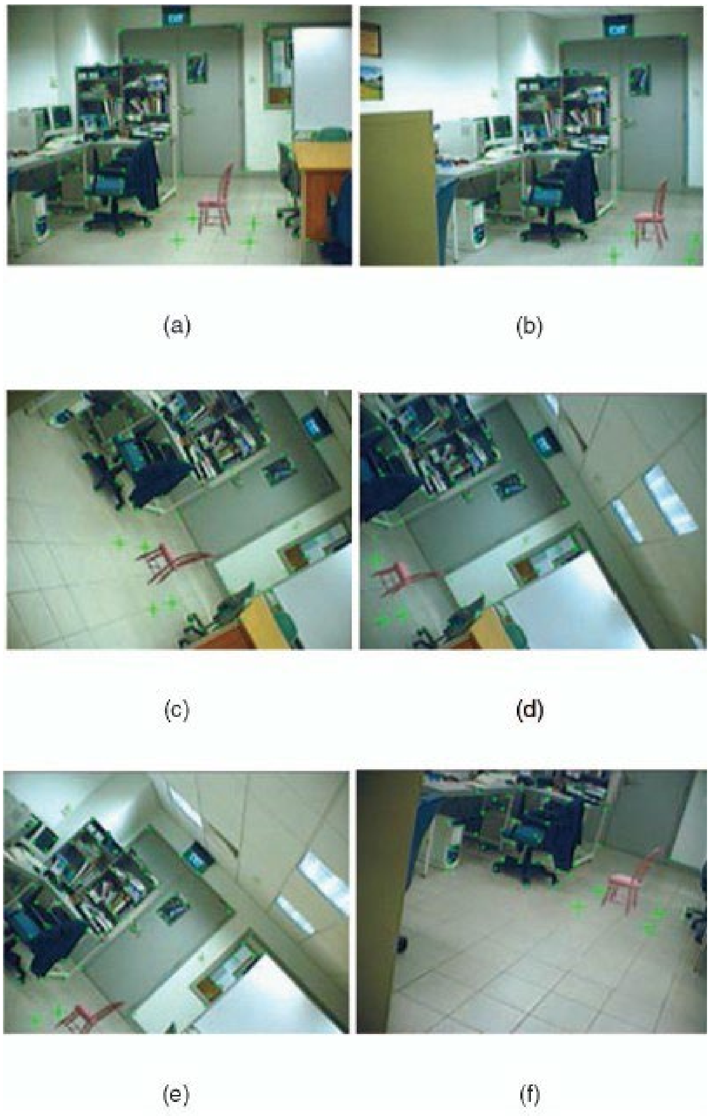
(a) (b) (c) (d) (e) (f)

Figure 1 – Experiment 1

Figure 3 shows the RMS error between the tracked natural features and those projected using the original natural features as input data. Figure 3(a) and Figure 3(b) show the estimation errors for the first and second indoor experiments, respectively. In the figure, the black curve represents the error using the original data, while the blue curve represents the error using the standardized data, where the abscissa axis represents the image frame of the image sequence and the ordinate axis represents the error value in the pixel.

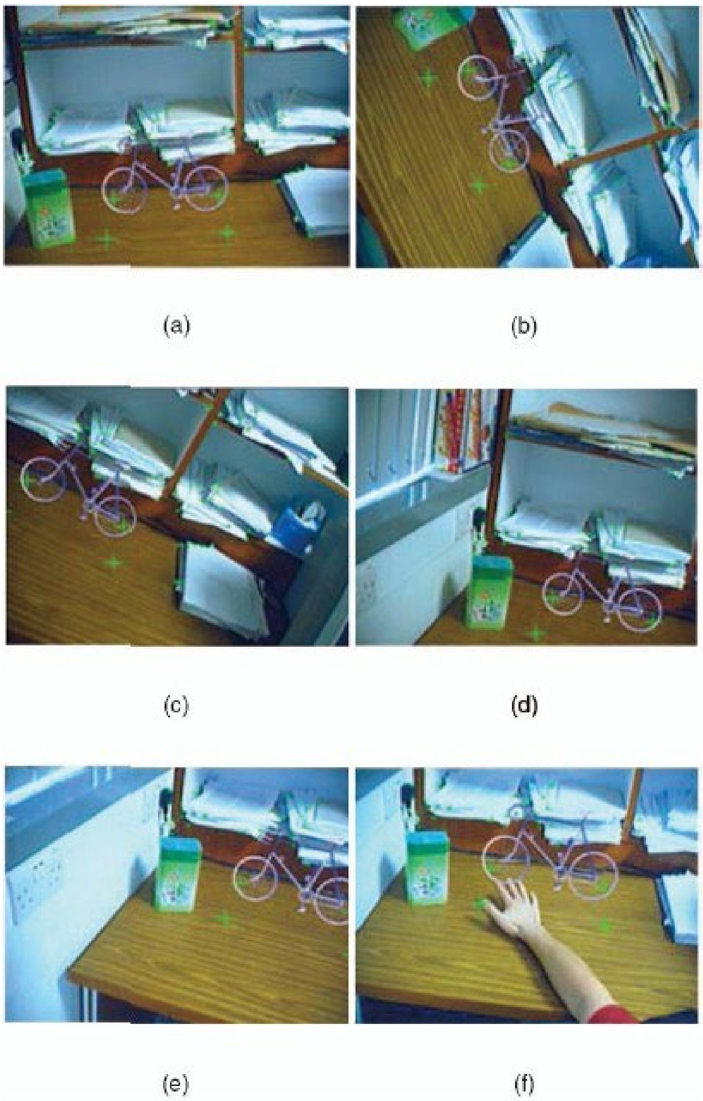

Figure 2 – Experiment 2

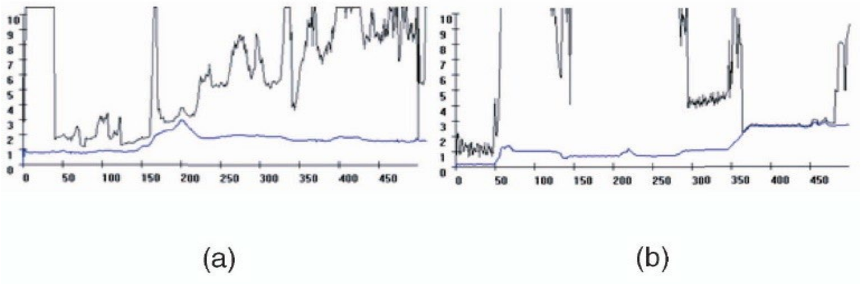

Figure 3 – Estimation error of experiment

As shown in Figure 3(a), the error of the normalized data is relatively stable when the camera moves and the pixel value is below 4.0, and the virtual object can be stably enhanced in the specified range.In a second indoor experiment, a virtual chair was placed on the real floor. As shown in Figure 1-2, the four points marked with "+" are transformed during the enhancement process. The enhancement with normalized natural features as input data is shown in Figure 3(b).

6. Comparison with other methods

The purpose of the experiments performed here is to compare the results obtained with the current method with those obtained with other methods. In this experiment, a sign with four corners is placed on the table, and then ARToolKit, KLT tracker, and the method described in this article are used to place a virtual object over the sign, as shown in Figure 4.

Consequently, the results of the first, second, and third lines in Figure 4 are the third, twenty-first, and forty-ninth frames in the video sequence and the first, second, and third frames in the video sequence. Figures 4(a1), 4(a2) and 4(a3) are images taken with ARToolKit, KLT, and the method described in this paper when the marker is on the table and the bicycle is on the marker. During the tracking process, if a small part of the marker is obscured by the hand, the virtual object cannot be placed on the marker using the ARToolKit method, as shown in Figure 4(a1), while in Figure 4(b2), if the marker is obscured by the hand at one of the four corners, the tracking using the KLT method is not correct, so the virtual bicycle is not placed on the marker correctly. In addition, Figures 4(c1) and 4(c2) show that if the part of the mark is outside the field of this video, the virtual bike cannot be placed on the mark with the artoolkit and KLT trackers, even if the mark is partially obscured or one of the four corners is tracked incorrectly, the virtual bike can be placed on the marker stably, these experiments show the advantages of the method described in this paper.

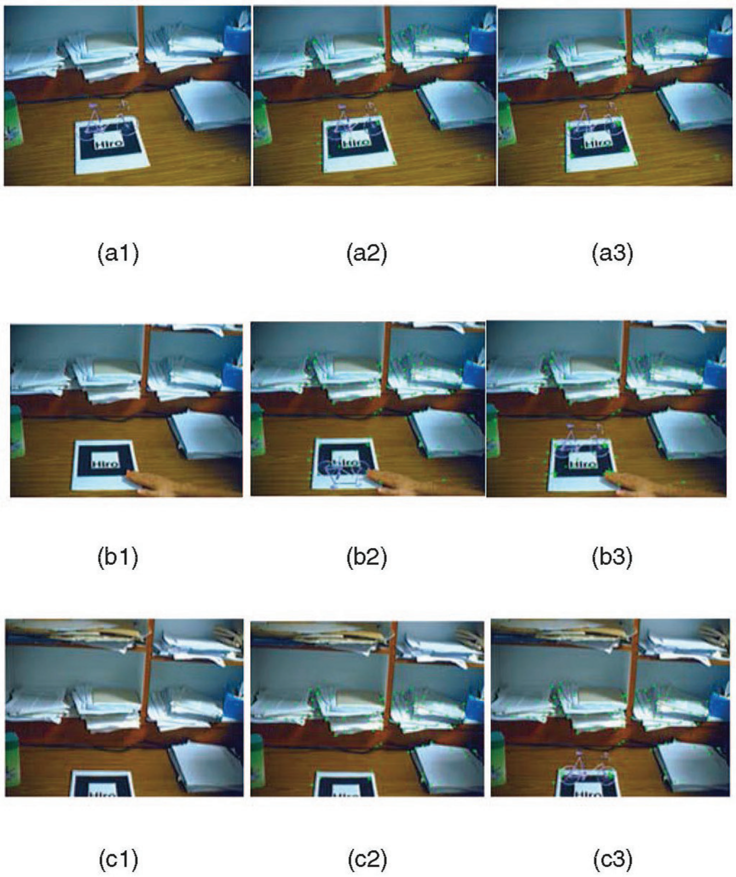

Figure 4 – Experiment 3

7. Conclusion

In this paper a registration method using natural features is presented, in this registration algorithm the classical KLT tracker is used to get natural features. The virtual objects can be superimposed in any region specified by the user and these points can be tracked by using projection reconstruction techniques. The biggest strength of this method is the virtual objects can be superimposed in the specified region even if some parts of the region are shaded during tracking, according to the user's requirements in any specified region in the natural environment.

References

1. M. Rosenthal, A. State, H. Lee, G. Hirota, J.Ackerman, K. Keller,E.D. Pisano, M. Jiroutek, K. Muller, and H. Fuchs, “Augmented Reality Guidance for Needle Biopsies: An Initial Randomized,Controlled Trial in Phantoms,” [M] Medical Image Analysis, vol. 6, no. 3,pp. 313-320, 2002.

2. S. Julier, M. Lanzagorta, Y. Baillot, L. Rosenblum, and S. Feiner,“Information Filtering for Mobile Augmented Reality,” [J] Proc. IEEE Int’l Symp. Augmented Reality, pp. 3-11, 2000.

3. R. Behringer, “Registration for Outdoor Augmented Reality Applications Using Computer Vision Techniques and Hybrid Sensors,” [M] Proc. Virtual Reality Ann. Int’l Symp., pp. 244-251, 1999.

4. R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B.MacIntyre, “Recent Advances inAugmented Reality,” [J]. IEEE Computer Graphics and Applications, vol. 21, no. 6, pp. 34-47, Nov.-Dec. 2001.

5. J. Shi and C. Tomasi, “Good Features to Track,” [J] Proc. IEEE Conf.Computer Vision and Pattern Recognition, pp. 593-600, 1994.

6. C. Tomasi and T. Kanade, “Detection and Tracking of Point Features,” Technical Report CMU-CS-91-132, Carnegie Mellon Univ., 1991.

7. G.D. Hager and P.N. Belhumeur, “Real-Time Tracking of Image Regions with Changes in Geometry and Illumination,” [J]Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 403-410, 1996.

8. B. Georgescu and P. Meer, “Point Matching under Larger Image Deformations and Illumination Changes,” [J]IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, pp. 674-689, 2004.

9. T. Tommasini, A. Fusiello, E. Trucco, and V. Roberto, “Making Good Features Track Better,” [J] Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, pp. 178-183, 1998.

10. U. Neumann and Y. Suya, “Natural Feature Tracking for 1, pp. 53-64, 1999.

11. V. Ferrari, T. Tuytelaars, and L. Van Gool, “Markerless Augmented Reality with a Real-Time Affine Region Tracker,” [J] Proc. IEEE and ACM Int’l Symp. Augmented Reality, pp. 87-96, 2001.

12. Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong, “A Robust Technique for Matching Two Uncalibrated Images through the Recovery of the Unknown Epipolar Geometry,” Artificial Intelligence J., vol. 78, pp. 87-119, 1995.

13. R.I. Hartley, “In Defense of the Eight-Point Algorithm,” [J] IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 6, pp. 580-593, June 1997.

14. H. Kato and M. Billinghurst, “Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System,” [J] Proc. Second IEEE and ACM Int’l Workshop Augmented Reality, pp. 85-94 Oct. 1999.

15. M.L. Yuan, S.K. Ong, and A.Y. C. Nee, “Registration Using Projective Reconstruction Technique for Augmented Reality Systems,” [J] IEEE Trans. Visualization and Computer Graphics, vol. 11,no. 3, pp. 254-264, May-June 2005.