Robust Pose Estimation for Outdoor Mixed Reality with Sensor Fusion

ZhiYing Zhou, Jayashree Karlekar, Daniel Hii, Miriam Schneider, Weiquan Lu, and Stephen Wittkopf

Interactive Multimedia Lab, Department of Electrical Computer Engineering, Department of Architecture, National University of Singapore, Singapore
{Elezzy,elejayas,akidhjc,eleschne,elelwq,akiskw}@nus.edu.sg

Abstract. We present a sensor fusion based technique for outdoor augmented reality system for mobile devices using GPS, gyroscope, and geo-referenced 3D models of the urban environment. Geo-spatial interaction not only provides overlays of the existing environment but compliments with other data such as location-specific photos, videos and other information from different time periods enhancing the overall user experience of augmented reality. To provide robust pose estimation of the camera relative to the world coordinates, firstly, GPS and gyroscope are used to obtain the rough estimation. Secondly, model based silhouette tracking and sensor fusion approach is used to refine the rough estimation and to provide seamless media rich augmentation of 3D textured models.

1 Introduction

Ubiquitous availability of high-end mobile devices has given rise to increased interest in mobile based applications. These devices now come with high-resolution digital cameras, displays, graphical capabilities and broadband connectivity which were available on desktop computers only few years ago. With these increased technological accessibility on move with geo-referencing provided by GPS (Global Positioning System), the user can exchange location-specific multimedia information (images, video, text etc.) anywhere anytime. With increased interest in graphical content creation for virtual reality, the user's location awareness and perception can be enhanced beyond multimedia by overlaying geo-referenced graphical contents. Mixed reality bridges the gap between real and virtual 3D world by exploiting the technologies and methods developed for virtual reality and the knowledge of the user location provided by GPS. Possible outdoor mixed reality application comprises architectural walkthroughs, tourism, exploration etc.

Mobile outdoor mixed reality overlays the 3D digital models having rich textures over a user's view of the real world. Augmentation of location specific information in graphical format in the user's view enhances the real world experience beyond normal. The user can change the time period of interest via the device. Wireless communication between mobile client and server allows users to share location-specific photos, videos and other information from different time periods. This illusion is

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2009, LNCS 5616, pp. 281–289, 2009.

possible if camera location and orientation in global space is accurately known. Lack of accuracy can cause complete failure of coexistence of real and virtual world. A successful mixed reality system must enhance situational awareness and should have the following attributes [13]:

- Runs interactively and in real time
- Combines real and virtual worlds in real environment
- Aligns real and virtual objects

With reference to above mentioned requirements, many approaches have been proposed in literature. These approaches can be divided into following three categories depending on the technology that is being used for estimating camera position and orientation [5]:

- External tracking devices like GPS and inertial sensors
- Vision approaches based on camera data alone
- Hybrid approaches to overcome the drawbacks of any single sensing solution

Inertial sensors provide estimate of camera pose at high-sampling rates, albeit at the cost of accuracy. Precision of these devices is less than desired for true visual merging for AR systems. Vision based approaches reply solely on camera data to estimate camera pose. These systems normally employ feature based detection and tracking. Camera based tracking generally provides the best estimate under reasonable viewing conditions e.g. small motion. However, their performance deteriorates under large viewing changes due to motion blur, noise, inaccurate tracking and high computational cost. Recently, hybrid approaches are proposed which try to combine strengths of each individual approach to compensate others limitation. These systems utilize data of inertial sensors as a rough estimate of camera pose which vision system refines further.

The paper describes one such hybrid approach for outdoor mixed reality applications for handheld devices. The proposed framework uses data from GPS, gyroscope and camera attached to it for final pose estimation. The system comprises the following modules:

- Multi-sensor fusion of location and orientation data (Section1.2 System Overview)
- Digital content creation and 3D CAD modeling (Section 2 3D Textured Modeling)
- Model based silhouette tracking for registering real and virtual objects (Section 3 Robust Geo-referencing)
- Image and 3D model rendering on mobile device (Section 4 Implementation on Mobile Device)
- Discussion and conclusion (Section 5)

1.1 Related Work

Marker based approaches provide a robust and stable solution for prepared environments. However, is it infeasible to equip large outdoor spaces with fiducial markers. Approaches presented in [1], [3], and [4] provide examples of outdoor markerless augmented reality by using GPS data with/without inertial data. GPS combined with

3D gyroscope data provides the fast and rough estimate of camera pose in unprepared environments. However, accuracy of these devices in urban environment is a matter of great concern. Computer vision techniques also estimate the 3D pose from the camera data, however, at the cost of increased computational requirement which is of great concern given the limited resources available on the mobile devices. To overcome the practical limitations of these different modalities in the context of mobile devices, hybrid approaches are normally employed for estimating the 6 DOF (degrees of freedom) of the camera [5], [6], [7], and [10]. These approaches do the sensor fusion by estimating rough camera pose from GPS and gyroscope data while vision techniques are employed for the refinement purpose only. Refinement in most cases is obtained by first detecting and matching visual features and recovery of camera parameters from those matches. These vision techniques try to minimize the drifts associated with GPS and inertial sensors for seamless overlaying of graphical data.

Model based vision techniques are the most preferred choice for the registration of real and virtual worlds. Computationally heavy line-tracking approach is presented in [7] whereas edge tracking from outlines of objects is proposed in [6]. Ref. [9] provides the comprehensive review of literature for model-based tracking of rigid objects. However, such solutions are computationally intensive, especially when dealing with the occlusions.

1.2 System Overview

In this paper, we propose the model based silhouette tracking approach to minimize the drifts in measurements of GPS and gyroscope data. The silhouette tracking approach is easy to extract, track and implement as compared to other techniques proposed in literature. This approach automatically extracts significant edges which are appropriate for tracking. Mask based moving edges approach presented in [8] is used in this work for edge tracking. The approach assumes that solid-textured 3D graphical models are available as opposed to wireframe models. Silhouette of the rendered data is extracted from solid models only as they provide automatic culling of occluding edges. Overall sensor fusion approach is illustrated in Fig. 1.

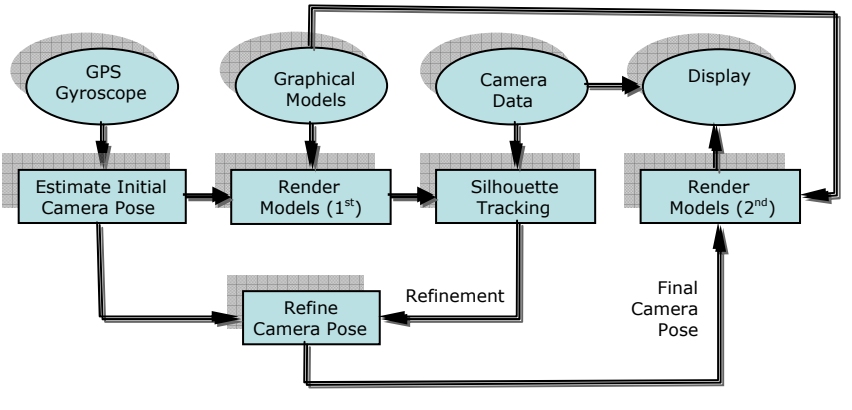


Fig. 1. System Overview

GPS and gyroscope provides initial rough estimate of the camera location and orientation. To remove jitter, the data from gyroscope is subjected to filtering before it is used. This initial camera pose is used to render the graphical models (1st block). The misalignment between rendered models and real data acquired from camera is estimated by silhouette tracking module. Final camera pose is estimated by combining initial pose and refinement obtained from silhouette tracking which is used to re-render the 3D graphical data (2nd block) for final overlaying on real data which is then displayed to the user. The realignment takes place for every frame of the video sequence.

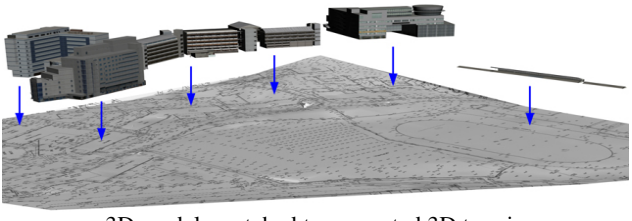
2 3D Textured Modeling

In order to augment past or future media with user's real environment, virtual 3D models of the urban environment having rich content are required for successful augmented reality applications. Existing approaches use 3D CAD modeling technique based on terrain maps. We have supplemented these maps with innovative 3D image based modeling technologies, whereby corresponding reference points across a series of images automatically create a 3D model textures on coarse geometrical models for visually pleasing effects. We obtained required logistics from G-element [11] for 3D content creation. These models are created in virtual space with geo-referenced origin. Fine alignment of altitude/height of virtual and physical worlds based on terrain map is done manually. Geo-specific and foot-print precise façade alignment is illustrated in Fig. 2.

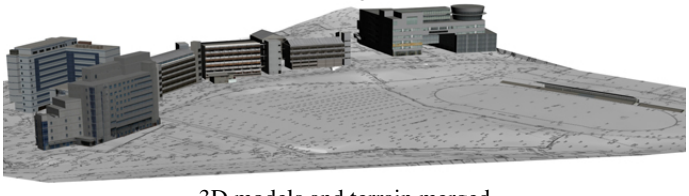
3 Robust Geo-referencing

For successful seamless augmentation, accurate and robust estimation of camera position and orientation in geo-referenced user frame is the first and foremost requirement. The systems normally employ GPS and gyroscope for position measurements and orientation respectively. Given the resolution and accuracy of these devices, they provide the rough estimate of camera parameters only. Precision of these devices is not satisfactory for mixed reality applications. Vision based approach to estimate the correct camera parameters are normally employed. However, models based approach is by far the preferred choice.

Feature matching and tracking is not suitable for model based approaches as feature descriptors associated with rendered model are *artificial* whereas those associated with camera data are *natural* ones. The mismatching between them is normally caused due to illumination changes, inaccuracies arising from texture acquisition and mapping process etc. Model based edge tracking is robust as feature involved are edges which are detected under varying illumination changes and fast as matching is performed in one-direction only, i.e. the direction perpendicular to edge orientation (motion along the edge is not perceived due to aperture problem). We propose silhouette tracking approach as edges corresponding to model and real image gets cluttered due to large viewing distances which renders the edge tracking approach impractical for AR applications. Camera based tracking is employed as a fallback mechanism whenever silhouette tracking fails due to unavailability of clear outline, too small viewing distances etc.



3D models matched to generated 3D terrain



3D models and terrain merged

Fig. 2. Fine alignment of virtual worlds with 3D terrain data

3.1 Silhouette Tracking

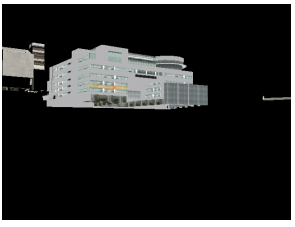
The approach presented here is free of cluttered edges, edges caused due to presence of elements/subjects in lower parts of the video data which is typical to urban environment. Silhouette tracking is performed on grayscale images and overall approach is presented in tabular format in Fig. 3. The approach assumes the presence of well-structured sky-line within the field of view of the camera. The resolution of images is 320 X 240 pixels.

First row illustrates the video image and virtual image rendered using approximate camera pose obtained from GPS and gyroscope. Edge extraction is performed using Canny edge detector on the grayscale version of these images (second row) obtained from camera feed and rendered data. As mentioned earlier, due to large viewing distances edges get cluttered making them unsuitable for tracking without further processing which is obvious from these images. The approach presented in [7] detects and tracks lines as opposed to edges to cull the weak edges. However, detection of lines is complex and tracking could fail because of occlusion. The approach presented in [6] detects outlines of rendered models for tracking purposes.

Extraction and tracking of outline is difficult in urban scenarios due to high density of objects, presence of other objects/subjects such as trees, traffic, pedestrians etc. We resolved these issues by using silhouette tracking (third row). As mentioned, silhouettes are free of cluttered edges, easy to extract and track as opposed to previous approaches. Tracking is performed using moving edges algorithm proposed in [2], [8]. The tracking results obtained from edge matching are presented in last row of Fig. 3. From these correspondences, along with the depth values obtained from Z-buffer, 6-DOF camera parameters are obtained by assuming perspective camera projection model.

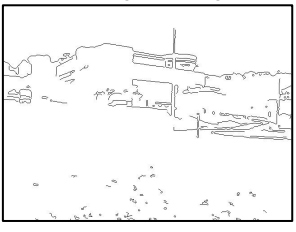
3.2 Conversion of GPS Data to Local Tangent Plane

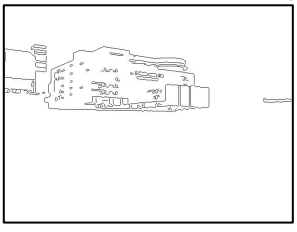
The data from GPS is in Earth centered-Earth fixed geodetic (ECEF-g) coordinate system. The geodetic coordinates of this frame are usually written as $<\lambda, \phi$, h> for



Original Image

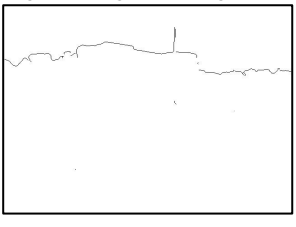
Rendered virtual scene before alignment





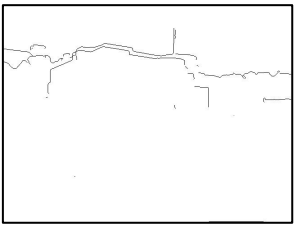
Original image after edge detection

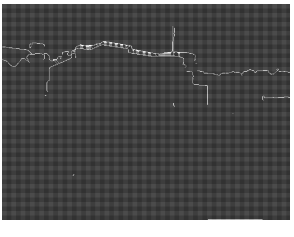
Rendered scene after edge detection



Silhouette of the original image

Silhouette of the virtual scene





Overlaid silhouettes

Correspondences obtained from edge tracking

Fig. 3. Illustration of silhouette tracking for correct estimation of camera parameters

latitude, longitude, and altitude (height) respectively. More convenient reference frame is ECEF rectangular (ECEF-r) system. The ECEF-r coordinates can be obtained from geodetic as:

$$x = (h+N)\cos\lambda\cos\phi$$
$$y = (h+N)\cos\lambda\sin\phi$$
$$z = (h+(1-e^2)N)\sin\lambda$$

These coordinates in ECEF-r frame are further transformed in Local Tangent Plane (LTP). This is an orthogonal, rectangular, reference system defined with its origin at an arbitrary point on the Earth's surface. The transformation from ECEF-r to LTP is done using the following formula:

$$x_t = \begin{bmatrix} e \\ n \\ u \end{bmatrix} = \begin{bmatrix} -\sin\phi & \cos\phi & 0 \\ -\cos\phi\sin\lambda & -\sin\lambda\sin\phi & \cos\lambda \\ \cos\lambda\cos\phi & \cos\lambda\sin\phi & \sin\lambda \end{bmatrix} \begin{bmatrix} x-x_0 \\ y-y_0 \\ z-z_0 \end{bmatrix}$$

where $<x_o, y_o, z_o>$ is the origin of LTP expressed in ECEF-r coordinate system [12]. We have taken this arbitrary point to coincide with an origin used for 3D graphical modeling.

4 Implementation on Mobile Device

4.1 System Specifications

The main hardware used in the research consist of the single camera PDA (personal digital assistant) (HP iPAQ rw6828 Multimedia Messenger), the gyroscope (Vitec 3D Sensor TDS01V) and the GPS device (HOLUX M-1000) as shown in Figure 4. The iPAQ PDA is running on Windows Mobile 5 and generates a video at resolution of 320 X 240 pixels at 15 fps (frames per second). The gyroscope connects using USB port while the GPS device connects using Bluetooth.

Software development platform used is C++ whereas graphic models are rendered with DirectX SDK.

4.2 Implementation

The application is simulated on UMPC having a database of relevant 3D models stored on it. The application provides augmentation of 3D textured models of urban data from geospatial databases. The time stamps associated with each object allows rendering of them accordingly to provide historical perspective of the specific location. Fig. 5 illustrates the concept of augmented reality in which rendered model is overlaid on the real data using camera position obtained from hybrid-sensor fusion approach presented in this paper. Porting the solution to mobile is under development.

The Department of Architecture at the National University of Singapore has large collection of historic photographs of Singapore. We have selected a sizeable subset of these photographs describing an area of interest, annotated them with geo-location metadata as a basis for the virtual models. This facilitates the exploration of the concept of mediated *'time travel'* by which the explorer can peer through the mist of time through a hand-held augmented reality device.

Fig. 4. The gyroscope, the GPS device and PDA

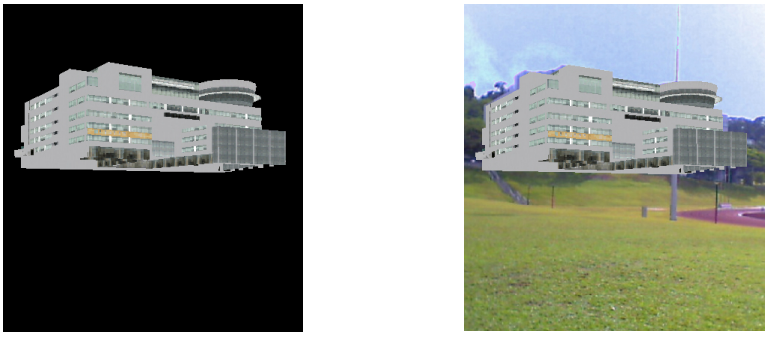


Fig. 5. Rendered graphical model using refined camera pose (left), overlaying of 3D texture model on captured image (right)

5 Conclusion and Discussion

In this paper we have proposed the hybrid sensor fusion approach for outdoor augmented reality applications. The proposed approach offers following advantages over the existing approaches:

- Silhouette extraction is robust under varying lighting conditions
- Tracking of silhouette is fast as search is carried out in direction normal to the edge orientation as opposed to two-dimensional search used in feature tracking
- Robust to occlusion which is common in outdoor environments

However, situations can arise where extraction of silhouette may not be possible and algorithm fails. One such case can occur when user captures the video from very near distances, in which case the scene occupies the whole field of view and silhouette tracking algorithm may not work due to unavailability of clear sky-line. However, in this scenario vision based alignment may not be needed as misalignment is not perceived and geo-referencing is still possible with the help of GPS and gyroscope data. In another situation outline may not be detectable due to poor contrast. In this case camera based tracking is employed as a fall back mechanism for camera parameter estimation.

Acknowledgements. The project is funded by Singapore A*Star Project No. 062-130-0054 (WBS R-263-000-458-305): i-Explore Interactive Exploration of Cityscapes through Space and Time.

References

1. Azuma, R., Lee, J.W., Jiang, B., Park, J., You, S., Neumann, U.: Tracking in Unprepared Environments for Augmented Reality Systems. Computers and Graphics 23(6), 787–793 (1999)
2. Bouthemy, P.: A Maximum Likelihood Framework for Determining Moving Edges. IEEE Trans. Pattern Analysis and Machine Intelligence 11, 499–511 (1989)
3. Feiner, S., MacIntyre, B., Hollerer, T., Webster, A.: Touring Machine: Prototyping 3D Mobile Augmented Reality Systems for Exploring the Urban Environment. In: Proc. ISWC 1997, pp. 74–81 (1997)
4. Honkamaa, P., Siltanen, S., Jappinen, J., Woodward, C., Korkalo., O.: Interactive Outdoor Mobile Augmentation using Markerless Tracking and GPS. In: Proc. International Conference on Virtual Reality (2007)
5. Hu., Z., Uchimura, K.: Fusion of Vision, GPS and 3D Gyro Data in Solving Camera Registration Problem for Direct Visual Navigation. Int. Journal of ITS Research 4(1) (2006)
6. Reitmayr, G., Drummond, T.W.: Going Out: Robust Model-based Tracking for Outdoor Augmented Reality. In: IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 109–118 (2006)
7. Jiang, B., You., S., Neumann, U.: A Robust Hybrid Tracking System for Outdoor Augmented Reality. In: IEEE Proc. Virtual Reality (2004)
8. Karlekar, J., Le, S.N., Fang, A.: Tracking of Articulated Pose and Motion with a Markerized Grid Suit. In: Proc. Int. Conf. on Pattern Recognition ICPR 2008 (2008)
9. Lepetit, V., Fua, P.: Monocular Model-based 3D Tracking of Rigid Objects: A Survey. Foundations and Trends in Computer Graphics and Vision 1, 1–89 (2005)
10. You., S., Neumann, U.: Fusion of Vision and Gyro Tracking for Robust Augmented Reality Registration. In: IEEE Proc. Virtual Reality (2001)
11. G-element, `http://www.gelement.com/`
12. Farrell, J., Barth, M.: The Global Positioning System and Inertial Navigation. McGraw-Hill, New York (1999)
13. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent Advances in Augmented Reality. IEEE Computer Graphics and Applications (November/December 2001)