MODEL-BASED LOCALIZATION AND DRIFT-FREE USER TRACKING FOR OUTDOOR AUGMENTED REALITY

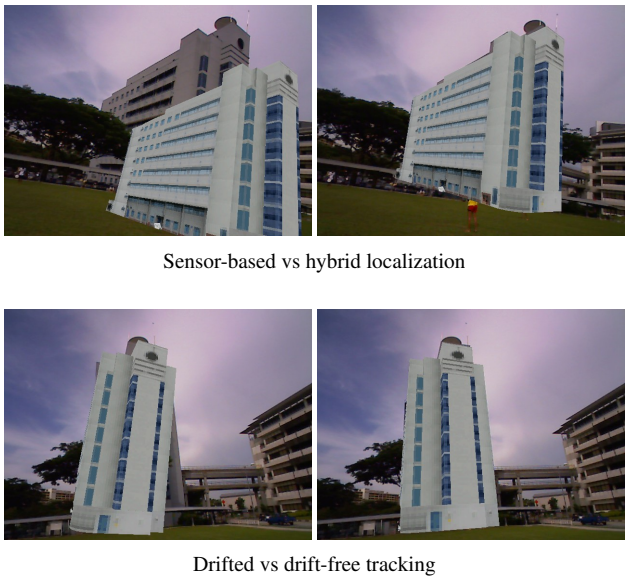
Jayashree Karlekar, Steven ZhiYing Zhou, Yuta Nakayama, Weiquan Lu, Zhi Chang Loh, Daniel Hii*

Interactive Multimedia Lab., Dept. of ECE
National University of Singapore
Email: *elezzy@nus.edu.sg

ABSTRACT

In this paper we present a novel model-based hybrid technique for user localization and drift-free tracking in urban environments. In outdoor augmented reality, instantaneous 6-DoF user localization is achieved with position and orientation sensors such as GPS and gyroscopes. Initial pose obtained with these sensors is dominated by large positional errors due to coarse granularity of GPS data. Subsequent tracking is also erroneous as gyroscopes are prone to drifts and often need recalibration. We propose to use model-to-image registration technique to refine initial rough estimate for accurate user localization. Large positional errors in user localization are mitigated by aligning silhouettes of the model with that of the camera image using shape context descriptors as they are invariant to translation, scale and rotational errors. Once initialized, drift-free tracking is achieved by combining frame-to-frame and model-to-frame feature tracking. Frame-to-frame tracking is done by matching corners whereas edges are used for model-to-frame silhouette tracking. Final camera pose is obtained with M-estimators.

***Keywords*—** Mobile Augmented Reality, Drift-free Camera tracking, User Localization, Shape Matching



Sensor-based vs hybrid localization

Drifted vs drift-free tracking

Fig. 1. **First Row:** GPS and inertial sensors together provide 6-DoF rough estimate of the camera in outdoor environments (left). Model-based shape matching technique to mitigate localization errors (right). **Second Row:** Error accumulation over time causes drifts in camera pose estimation (left). Combined region (frame-to-frame) and silhouette (model-to-frame) tracking can reduce tracking drifts (right).

1. INTRODUCTION

The evolution of mobile-computing, location sensing and wireless networking has created a new class of computing: *context-aware computing*. Mobile computing devices such as PDAs have access to information processing and communication capabilities but do not necessarily have any awareness of the context in which they operate. Context-aware computing describes the special capability of an information infrastructure to recognize and react to the real world context. The most critical aspect of context then is the location. One such context-aware technology is mobile augmented reality (MAR) which combines a users view of real world with location specific information. Such information could be in the form of simple text, image, multimedia or 3D graphics. Possible applications of MAR comprise architectural walkthroughs, tourism, exploration etc.

The most important aspect of MAR system is to identify the location and orientation of the user to retrieve the context so as to present context-aware information thereby enhancing the general awareness of the surrounding. Accurate estimation of camera pose in global space is the most important aspect to provide such mixed illusion. Lack of accuracy can cause complete failure of coexistence of real and virtual worlds. In MAR systems, instantaneous 6-DoF user localization is achieved with position and orientation sensors such as GPS and gyroscopes. Initial pose obtained with these sensors is dominated by large positional errors due to coarse granularity of GPS data. Subsequent tracking is also erroneous as gyroscopes are prone to drifts and often need recalibration. In some applications accuracy of these devices may be adequate whereas in others it is

The work is financially supported by Singapore A*Star Project 062-130-0054 (WBS R-263-000-458-305): i-Explore Interactive Exploration of Cityscapes through Space and Time. The project is also in collaboration with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University.

less than desired for true visual merging. Vision techniques can be employed to mitigate sensor errors for accurate user localization and tracking in urban environments as illustrated in Fig. 1.

To overcome the practical limitations of different modalities in the context of outdoor environments, hybrid approaches are normally employed. These approaches utilize data of inertial sensors as a rough estimate of the camera pose while vision technique refines it further. This paper presents one such novel hybrid model-based approach for outdoor mixed reality applications in which automatic initialization and drift-free tracking is achieved. Automatic initialization is done by refining the rough camera pose obtained from sensors by matching silhouettes of model data and image data with shape context descriptors [1]. These descriptors are invariant to translation, scale and rotational errors, which is very desirable for automatic alignment as initial pose estimates are often dominated by position errors due to coarse granularity of GPS data.

Once aligned, extended Kalman filter (EKF) of constant velocity model is initialized and system switches to tracking mode. Drift-free camera tracking is obtained by combining frame-to-frame and model-to-frame tracking. Frame-to-frame tracking is performed by using Harris corner detector and matching [2] whereas model-to-frame silhouette tracking is performed by moving edges algorithm of [3]. Model-to-frame matches serve as a feedback mechanism to correct the drifts in camera tracking. These different matches are then appropriately weighted with camera pose predicted from EKF. Linear system of equations is then solved for camera pose by iterative, reweighted least-squares.

2. RELATED WORK

Many approaches ranging from sensor-based to pure vision to hybrid have been proposed for outdoor augmented reality applications.

Pure vision based techniques such as [4, 5, 6] do image based object recognition to localize and track users in outdoor environments. User context here is acquired by querying the image database of labeled objects. On the other hand, hybrid approaches fuse data from different sensors such as camera, GPS, gyroscopes etc. In these approaches 3D georeferenced graphical models of the target surrounding serve as a context. These systems utilize data of inertial sensors as a rough pose estimate while vision system refines it further. Hybrid tracking approaches mainly differ based on which natural features are used by vision technique for tracking purposes. Approaches presented in [7, 8] use lines, [9, 10] uses edges while [11, 12] uses corners for tracking. These approaches take care of tracking only and initialization is often assumed known.

Early work in vision based automatic initialization is reported in [13]. The approach does model based automatic landmark detection and matching for rotational errors only. Approach presented in [14] does successive approximation of GPS data and edge tracking to converge to the correct pose.

3. CAMERA POSE ESTIMATION

Camera pose is estimated by 3D-2D correspondences and twists. A 3D point $P_w = (X_w, Y_w, Z_w, 1)^T$ represented by homogeneous coordinates in world frame is projected into point P_c in camera frame as:

$$P_C = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} P_w = GP_w$$

where, R (rotation) and T (translation) having 3 DoF each denote the relative transformation between world and camera coordinate frames. This rigid body motion is represented by $G \in SE(3)$ in homogenous coordinates as above. For every G there exists a twist $\hat{\xi} \in se(3)$, a 4×4 matrix with upper 3×3 component as a skew-symmetric matrix. Coordinates of twist are given by a 6-dimensional vector $\xi \in R^6$. G can be obtained from twist with exponential mapping: $G = e^{\hat{\xi}}$ [15].

Under camera motion, a point P_w^t at instance t is related to a point P_w^{t+1} at instance $t+1$ by rigid motion G:

$$P_w^{t+1} = GP_w^t = e^{\hat{\xi}} P_w^t \approx (I + \hat{\xi}) P_w^t.$$

After retaining first order terms only, their projection in image plane expressed in terms of twist is:

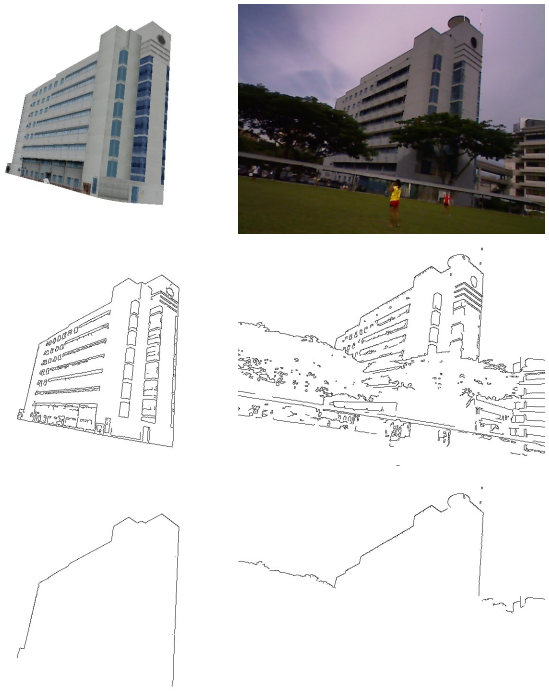
$$p^{t+1} - p^t = \begin{bmatrix} u_x \\ u_y \end{bmatrix} = J\xi. \quad (1)$$

Left hand side is the perceived optical flow **u** in image plane due to motion of camera and Jacobian J is a 2×6 matrix that relates these measurements to camera pose. For N feature matches, N equations of the above form are obtained. Then twist ξ that minimizes the sum of square residual errors $r_i = J\xi - u$ is obtained by solving following system of linear equations:

$$\xi = \arg\min_{\xi} \sum_i r_i^2 = \arg\min_{\xi} \|\mathbf{J}\xi - \mathbf{u}\|^2. \quad (2)$$

4. MODEL-BASED LOCALIZATION WITH SHAPE CONTEXT DESCRIPTOR

We propose a novel model-based non-iterative hybrid approach for automatic user localization in urban environments. Approach relies on model-to-frame features matching. Region based model-to-frame matching with features such as Harris [2], SIFT [16], or Fern [17] are not suitable as possible changes in illumination, inaccurate texture stitching etc. could lead to false matching and alignment. In outdoor environments, edges also tend to get cluttered due to large viewing distances making them unsuitable as potential feature candidate (Fig. 2). Reliable alignment is then achieved by matching appearance of the model outline to the image. Outline/silhouette from the model and image are then extracted as illustrated in Fig. 2. Without loss of generality, we assume textureless, georeferenced models of the environment are available. Given resource scarcity on mobile devices, such models are more desirable as they take less disk space and are fast to render.



Rendered model Camera Image

Fig. 2. First row: Rendered model and camera image. Illumination changes and errors in texture stitching are clearly visible. Second row: Edge detection results on input images. Third row: Contour extraction from rendered model and camera data for alignment.

With limited feature choices, model-based hybrid initialization with low level features is extremely challenging. Iterative closest point (ICP) [18] and distance transform (DT) based registrations have tendency to get stuck in local minima and hence inappropriate for localization due to large initial sensors errors.

Other approach for contour matching could be to prototype the appearance of the contour obtained from rendered model and match it in the camera data. One such prototyping of contours is proposed in [1] by *shape context* descriptor. The descriptor characterizes a particular contour point with respect to all other contour points. These points are randomly sampled from the contour shape. Relative distances, angles and normalization makes shape context descriptor invariant to translation, rotation and scale respectively. Once model and image contours are parameterized with shape descriptors, one-to-one sample point correspondences are established.

Assume that the set, $P = \{p_1, \ldots, p_n\}$, $p_i \in R^2$, of n points represent the points on shape obtained from model. Similarly, the set $Q = \{q_1, \ldots, q_m\}$, $q_i \in R^2$, of m samples represent the points on shape obtained from image (Fig. 3(a)). For each point p_i from model contour the best possible matching point q_j from the image contour is obtained. The local cost of matching a point p_i to a point q_j is $C_{ij} = C(p_i, q_j)$. As shape contexts are distributions represented as histograms, the matching cost is obtained with χ^2 test statistic.

Shapes are matched by minimizing cost for each point p_i on

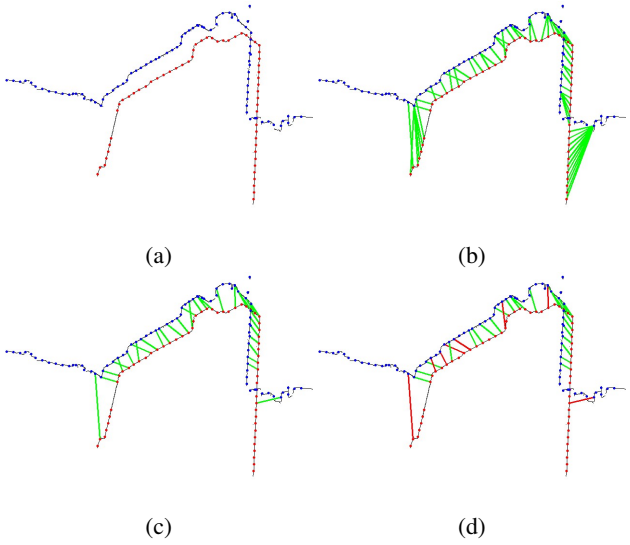


Fig. 3. Silhouette matching with *shape context* descriptor. (a) Sampling of model (red) and image (blue) contours with points. Image contour has extra points due to clutter of trees, unmodeled antenna on top of the building etc. Correspondences obtained with (b) forward and (c) bidirectional tracking. (d) Removing outliers (marked red) from (c) with figural continuity constraint.

model contour P to point q_j on image contour Q as:

$$C(P,Q) = \sum_{i=1}^{n} \min_{j} \; C(p_i, q_j) \tag{3}$$

We define such tracking as forward matching and results are demonstrated in Fig. 3(b). As the number of silhouette points n of model contour and m of image contour could be different, undesirable effects such as many-to-one matches are obtained. To obtain unique correspondences, we employ the bidirectional tracking in which matching cost of model-to-image (forward tracking) and image-to-model (backward tracking) shape matching costs combined together as:

$$C_{BT}(P,Q) = \frac{1}{2}[C(P,Q) + C(Q,P)] \tag{4}$$

Results of bidirectional tracking are illustrated in Fig. 3(c). Bidirectional tracking with shape context descriptor alone is not sufficient for reliable correspondences, especially in situations where extra clutter and/or partial shape data are encountered. This is normally the case for outdoor augmented reality as non-modeled elements such as pedestrians, vehicles, lamp-posts, trees etc. present in camera data give rise to extra clutter as well as occluding the desired target. Fig. 3(c) demonstrates the effect of these on shape context based matching. Extreme model shape points are getting matched to the points from clutter.

Robustness of shape context matching is then increased by introducing a figural continuity constraint proposed in [19]. The constraint states that, the two neighboring points on the model shape P should match to nearby points on the target shape Q.

Fig. 4. Combined tracking in progress. Extracted silhouette from image is overlaid by white color. Edge correspondences obtained from model-to-frame tracking are illustrated with red color while green ones coresponds to frame-to-frame region based corner matching.

However, this constraint needs points on the model shape to be ordered. Once ordering of points p_i is done using chain codes, neighboring matches are subjected to figural continuity constraint. Point pairs not obeying the constraint are examined and point correspondence having more shape context error from the pair is detected as outlier and excluded from pose calculations. The result of figural continuity constraint is illustrated in Fig. 3(d). Camera pose is obtained from remaining correct matches by iterative least squares using Eq. 2.

5. DRIFT-FREE CAMERA TRACKING

After correcting the initial camera pose, the system switches to tracking mode. EKF is initialized and updated at each frame which serves two purposes: one to provide estimate of the camera pose for next frame and another to smooth out the jittering effect of tracking. Drifts in tracking is a common problem, which arises due to inaccurate camera calibration, occlusion, incomplete and partial data etc. To achieve drift-free and robust tracking for long sequences, combined tracking of corners and edges is used. Frame-to-frame corner tracking captures the major camera motion while model-to-frame edge tracking provides feedback to correct drifts. Such combined tracking has been successfully used for tracking articulated [20] and complex rigid [21] objects. Approach proposed by [22] fuses corner and edge tracking in a unified manner.

Inspired by their results, we propose to use combined feature tracking for camera pose estimation in outdoor scenarios. Outdoor environment poses different challenges than controlled lab environments. Most of the time meaningful edges are difficult to extract for feasible model-to-frame matching as illustrated in Fig. 2. To provide resonable feedback to avoid drift in overall tracking, silhouette of the model is matched to the partial silhouette obtained from camera data using moving edges algorithm of [3]. Region based frame-to-frame tracking is carried out using Harris corner matching. Fig. 4 illustrates combined tracking process.

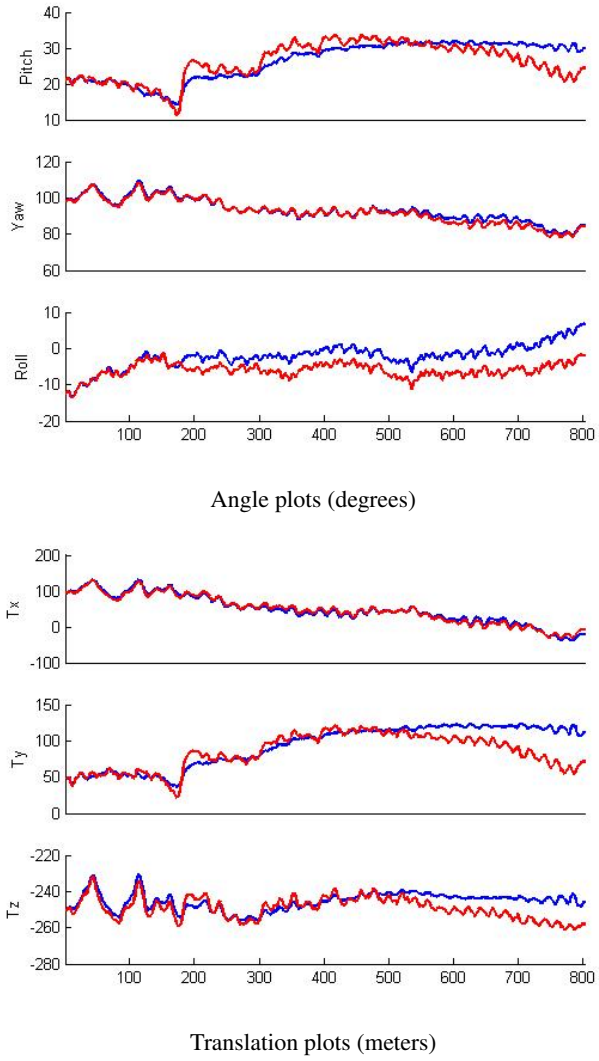
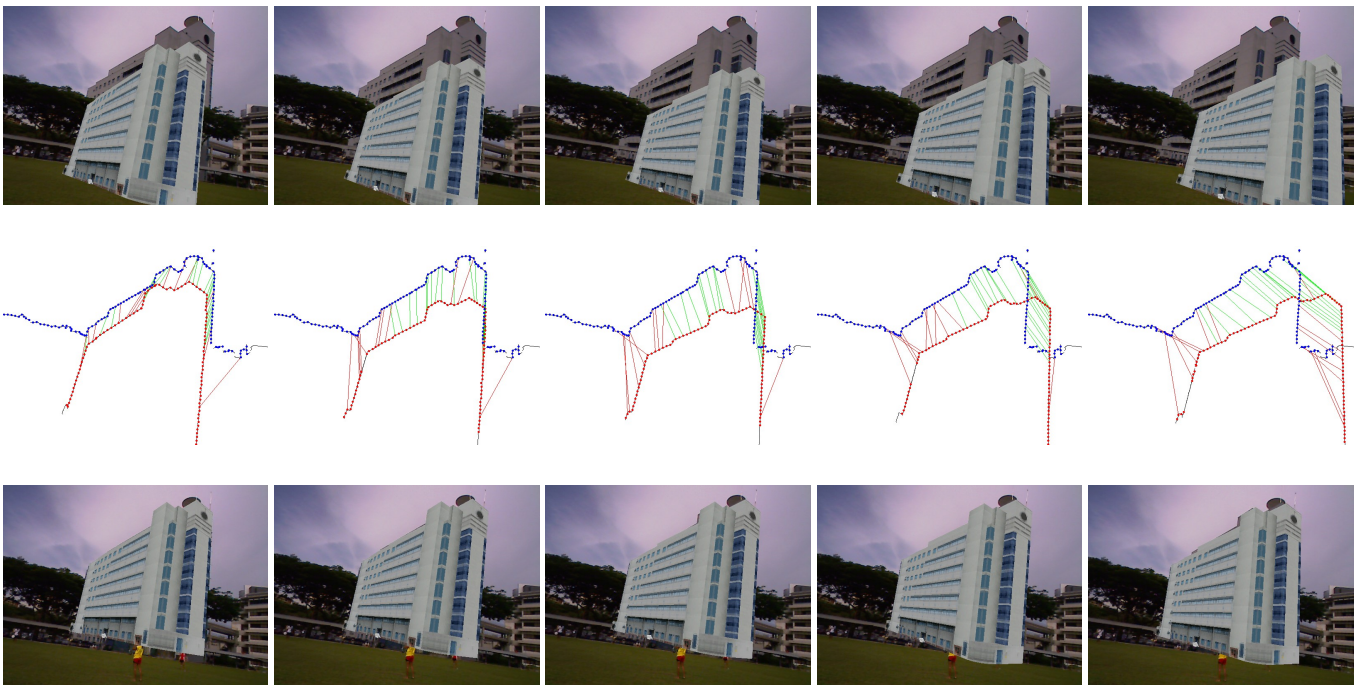


Fig. 5. Illustration of drifted (red) and drift-free (blue) tracking. Tracking based on corners only start drifting around 150 frame and which accumulates further.

Region based corner matching suffers from occlusion and mismatches arising due to repetitive structures present in man made environments. Similarly, edge tracking suffers from aperture problem. The moving edges algorithm captures motion only in direction perpendicular to edge while motion along the edge is not detected. To reduce the influence of different feature tracking algorithms, matches are appropriately weighted. Weighting is done using M-estimators. Camera pose at the current frame is predicted from EKF and Huber estimator [23] is used to generate weights. Final camera pose is predicted by iterative reweighted least-squares system of Eq. 2.



$C_x = 20.0m$	$C_x = 25.0m$	$C_x = 30.0m$	$C_x = 30.0m$	$C_x = 34.0m$
$C_y = 0.0m$	$C_y = 1.25m$	$C_y = 1.0m$	$C_y = 1.75m$	$C_y = 1.25m$
$C_z = 272.0m$	$C_z = 268.0m$	$C_z = 265.0m$	$C_z = 272.0m$	$C_z = 275.0m$

Fig. 6. User localization results with different GPS values. Absolute GPS coordinates are mapped to local tangent plane and are expressed in meters (C_x and C_z). C_y value reflects altitude (height), not captured by GPS, is varied from 0 to 1.75 meters for testing purposes. Orientation values are kept constant. First row illustrates initial pose estimate available from intertial sensors. Second row demonstrates shape matching and outlier detection step. Last row show the final alignment results.

6. RESULTS

The hardware system specification consists of the following devices. The Unibrain Fire-i camera with 1.9 mm lens is used for capturing the data at 320×240 resolution. Inertial sensor device used for orientation measurement is OS5000-S. The gyroscope is tightly coupled to the camera. The position sensor is HOLUX M-1000 GPS receiver. The system is targeted for the UMPC devices such as Raon Everun Note and runs at 9 frames per second. Graphical model consists of $1700+$ polygons. For illustration purpose only we are using textured model. Algorithm is equally applicable for models without textures. We demonstrate the tracking results for sequence consisting of $800+$ frames.

Results for automatic initialization with shape context based silhouette matching are presented in Fig. 6. Rough position estimate obtained from GPS have translation and scaling errors. Alignment results demonstrate the invariance of shape context based matching towards those errors. Accuracy of correct camera pose estimation entirely depends on the quality of matches between two shapes. Hence it is very essential to detect and reject outliers.

Once aligned, system switches to tracking mode. Results for corner tracking alone and combined tracking are presented in Fig. 5 and 7. Fig. 5 presents the quantitative error analysis for two methods. Corner based tracking starts drifting pretty early and never recovers. Marginal errors were observed in estimating yaw angle whereas large drifts (around 8-10 degrees) were resulted in pitch and roll angles. Similarly translation errors in all three directions were minimal till 500^{th} frame and large drifts in T_y, T_z are observed afterwards. Visual results of drift-free tracking are tabulated in Fig. 7 at different instances. First row present results of proposed tracking algorithm while second row shows results of corner tracking only. Visual results confirm the fact of quantitative error analysis. Drift-free tracking is achieved with combined tracking even at 801^{st} frame.

7. CONCLUSION

In this paper we have proposed hybrid non-iterative model-based user localization and drift-free tracking approach for outdoor mixed reality applications. Rough estimate of the camera pose available form position and orientation sensors is used to render the model. Silhouettes extracted from model and camera image are matched with shape context descriptor. Figural continuity constrains are imposed to get rid of outliers arising due to presence of clutter and partial data. The approach is non-iterative and avoids costly rendering of model at every iteration. Once initialized, algorithm switches to combined tracking

Frame 152 Frame 317 Frame 521 Frame 659 Frame 801

Fig. 7. Visual representation of camera tracking results at different instances. First row depicts tracking results with our proposed approach while second row is with traditional (based on corner tracking alone) approach. Drift keeps increasing by each passing frame.

mode. Different matches are appropriately weighted with respect to current pose predicted from EKF to handle short comings of individual trackers. Superior tracking results are obtained for longer video sequences. Combined drift-free tracking approach is even robust to slight initialization errors. Versatility of the proposed localization approach with shape context descriptor for different outdoor conditions is under consideration.

8. REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE. Trans. PAMI*, vol. 24, pp. 509–522, 2002.

[2] C. Harris and M. Stephens, "A combined corner and edge detection," *In. Proc. of The Fourth Alvey Vision Conference*, pp. 147–151, 1988.

[3] P. Bouthemy, "A maximum likelihood framework for determining moving edges," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 11, no. 5, pp. 499–511, 1989.

[4] I. Gordon and D.G. Lowe, "Scene modelling, recognition and tracking with invariant image features," *In Proc. ISMAR*, 2004.

[5] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," *In Proc. CVPR*, 2009.

[6] D.-N. Ta, W.-C. Chen, N. Gelfand, and K. Pulli, "Surftrac: Efficient tracking and continuous object recognition using local feature descriptors," *In Proc. CVPR*, 2009.

[7] B. Jiang, U. Neumann, and S. You, "A robust hybrid tracking system for outdoor augmented reality," *In Proc. Virtual Reality*, 2004.

[8] H. Wuest, F. Vial, and D. Stricker, "Adaptive line tracking with multiple hypotheses for augmented reality," *In Proc. ISMAR*, 2005.

[9] Z. Hu and K. Uchimura, "Fusion of vision, gps and 3d gyro data in solving camera registration problem for direct visual navigation," *Int. Journal of ITS Research*, vol. 4, no. 1, 2006.

[10] G. Reitmayr and T.W. Drummond, "Going out: Robust model-based tracking for outdoor augmented reality," *In Proc. ISMAR*, pp. 109–118, 2006.

[11] M. Aron, G. Simon, and M.-O. Berger, "Handling uncertain sensor data in vision-based camera tracking," *In Proc. ISMAR*, 2004.

[12] P. Honkamaa, S. Siltanen, J. Jappinen, C. Woodward, and O. Korkalo, "Interactive outdoor mobile augmentation using markerless tracking and gps," *In Proc. Virtual Reality*, 2007.

[13] K. Satoh, M. Anabuki, H. Yamamoto, and H. Tamura, "A hybrid registration method for outdoor augmented reality," *In Proc. ISAR*, 2001.

[14] G. Reitmayr and T.W. Drummond, "Initialisation for visual tracking in urban environments," *In Proc. ISMAR*, 2007.

[15] R. M. Murray, Z. Li, and S. S. Sastry, *A mathematical introduction to robotic manipulation*, CRC Press, 1994.

[16] D. Lowe, "Object recognition from local scale-invariant features," *In Proc. ICCV*, 1999.

[17] M. Ozuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," *In Proc. CVPR*, 2007.

[18] Z. Zhang, "Iterative point matching for registration of free form curves and surfaces," *Int. Journal of Computer Vision*, vol. 13, pp. 119–152, 1994.

[19] A. Thayananthan, B. Stenger, P.H.S. Torr, and R. Cipolla, "Shape context and chamfer matching in cluttered scenes," *In Proc. CVPR*, 2003.

[20] J. Gall, B. Rosenhahn, and H.-P. Seidel, "Drift-free tracking of rigid and articulated objects," *In Proc. CVPR*, 2008.

[21] E. Rosten and T.W. Drummond, "Fusing points and lines for high performance tracking," *In Proc. ICCV*, 2005.

[22] S. Birchfield and S.J. Pundlik, "Joint tracking of features and edges," *In Proc. CVPR*, 2008.

[23] Z. Zhang, "A tutorial with application to conic fitting," *Image and Vision Computing*, vol. 15, pp. 59–76, 1997.