A Literature Analysis of Object Tracking and Interactive Modeling in Videos for Augmented Reality

Reeba Mary Jacob, D. Narmadha

Department Of Information Technology
Karunya University
Karunya Nagar
Coimbatore -64114
India

Abstract

Considerable progress has been made in the offline tracking of 3D objects by processing the necessary information collected in real time. In contrast, the tracking of unknown 3D objects moving independently in an unknown environment has received less attention. This paper is an analysis of different approaches to track large number of 3D objects in real time and of the different methods to perform interactive modeling based on these tracked objects.

1. Introduction

Augmented Reality (AR) is the viewing of the real world environment with its elements augmented by computer generated sensory input. It finds uses in tracking technologies like digital camera, GPS , optical sensors and in applications like marketing and gaming. Implementation of Augmented Reality involves tracking of objects and the capability to interact with these objects in the enhanced environment.

This paper contains an analysis of the different tracking and detection methods for 3D objects in an unknown environment. An analysis is also provided on different interactive modeling processes which allow the user to manipulate the existing objects and add new objects to the environment being augmented. Section II lists the different tracking and mapping methods under consideration for analysis and Section III lists the interactive modeling techniques. Sections IV and V contain the detailed analysis of the methodologies and Section VI provides a comparison between the different methodologies in a tabular format.

2. Methodology For Detection And Tracking

Object tracking is an offline process which requires time and additional information. One methodology proposed by Klein G and Murray D in [1] is based on the monocular SLAM algorithm. The monocular SLAM algorithms are incremental mapping systems. It was initially used in robotic exploration. The same approach is applied to a hand-held camera. The SIFT based approach used by Kim K, Lepetit V and V Woow in [2] is a 3D tracking method capable of supporting several trained planar objects without affecting the real time performance. Another similar approach is the modified SIFT approach by Wagner D, Schmalstieg D, Bischof H in [3] which helped to create the first real-time six degree of freedom (6DOF) feature tracking system for mobile phones. The approach used by Pilet J, Saito H. in [4] combines tracking and image retrieval .The approach is a bi - layered cluster model which allows to index and retrieve the objects using tracks of features. The approach proposed by Kiyoung Kim , Vincent Lepetit and Woontack Woo in [5] performs the process of multiple 3D object tracking by combining the techniques of image retrieval and online Structure from Motion with both techniques running in parallel.

3. Methodology For Interactive Modelling

Interactive modeling can help in the reconstruction of the 3D models of real world objects. Pan Q, Reitmayr G, Drummond T in [6] proposes a system called ProFORMA which speeds up the on-line reconstruction of the object. The system uses AR for user's guidance in the reconstruction process. The modeling approach by Simon G. [7] allows the user to define the vertices for the object while moving the camera. The reconstruction and modeling is done by fitting 3D primitives to the 3D location. A similar approach is also used by Van den Hengel A, Hill R, Ward B and Dick A. in [8] which is combined with VideoTrace which is an off-line reconstruction technique for objects from videos.

4. Analysis Of Different Technique For Detection And Tracking

4.1. Analysis of SLAM algorithm in PTAM

The approach proposed by Georg Klein and David Murray in [1] is a Parallel Tracking And Mapping method also referred to as PTAM based approach .In this approach the estimating of camera pose is done based on the SLAM algorithm. The algorithm was developed for robotic exploration. The system is designed to track a hand - held camera within a small AR workspace. Mapping and tracking run as parallel thread. One thread performs tracking of the hand - held motion while the other thread produced a 3D map, of points features. It also involves batch optimization technique. The mapping is done based on keyframes which are processed using bundle adjustment. Initially a map is generated by the system which is a representation of the user's environment. The map contains a number of point features and keyframes which are snapshots from the hand - held camera taken at different points of time. When the point based tracking is completed, it is assumed that the map of 3D points has already been generated. The images from the camera are sent to the tracking system which maintains the real time pose estimates. Using these estimates the augmented graphics can be drawn on top of the frame. The coarsest scale features found are used in camera pose updating. To find a map point within the current frame a search is done around the point's predicted position in the image within a fixed range image. An 8X8 pixel patch search template is generated using which the best match within a fixed radius of the predicted location is found. The evaluation of the match is done using the zero - mean SSD score and FAST corner location within the predicted circular search region. The final camera pose is updated using both the coarse and fine sets of images measurement together. The quality of each frame is estimated based on the fraction of feature observation. If the fraction is below a certain threshold the quality of threshold is considered low and is not allowed to send new keyframes for the map. A recovery procedure is initiated when threshold is at most low.

In the mapping process the initial map is generated using the five point stereo algorithm and is expanded when new keyframes are added by the tracking system. The first keyframe is stored by the map and the 2D patch tracks are initialized. The five point stereo algorithm and RANSAC are used to generate the essential matrix and to triangulate the base map. The final map is refined using bundle adjustment. When adding new keyfames certain conditions have to be met. The tracking quality has to be good ; there should be a difference of atleast twenty frames from the last keyframe and the current keyframe ; and camera must have a minimum distance from nearest keypoint in the map. When no new keyframes are added the mapping thread has free time to improve the map. During the improvement if new measurements are found it is inserted into the map. It takes only a little effort to extract geometrics features, after the initial dominant plane is extracted as AR surface.

4.2. Analysis Of SIFT algorithm in Digilog Book

In the approach proposed by Kiyoung Kim , Vincent Lepetit and Woontack Woo in [2] it is assumed that one target is visible at a time. It is similar to the approach in [1] where the tracking and detection processes run as parallel threads. The foreground or the tracking thread tracks the feature from frame-to-frame and the background or the detection thread is used to recognize the targets and their pose.

First the SIFT features are extracted from all the references image of the target. A vocabulary tree is generated to quantize the features extracted . It also helps to recognize the targets quickly. A kd- tree for one tree per target is generated to store the SIFT features from the reference images of the target. The reference images are arranged based on the similarity with the input image. To find the match the SIFT features extracted from the input image is compared with those of the two candidate images which has highest similarity based on the kd-tree. The RANSAC algorithm is used to compute the similarity between the input and two reference images. When the detection module has recognized the target it sends the information about the pose to the tracking module. A 16X16 cross correlation window and bounded search region is used to match the features extracted in two images. Similar to the previous approach the Sum of the Square Difference (SSD) is computed between the patches of the current frame and the previous frame to improve the quality at every frame.

4.3. Analysis Of Guaranteed Framerates on Mobile Phones

The system proposed by Daniel Wagner, Dieter Schmalstieg and Horst Bischof in [3] is aimed at estimating the pose and tracking for real time in low-end devices like mobile phones. It is build on approaches described about in the modified SIFT[9] and Ferns[10] approaches to create a real time 6DOF feature tracking on mobile phones. The system aims to track and detect multiple targets for computationally

weak platforms. Originally the Difference of Gaussians (DOG) algorithm is used in SIFT for scale space search which detects the features and estimates the scale. As it is resource intensive, it is replaced with the FAST corner detector which has low computational requirements. To estimate the scale, the scale space is quantized to discrete steps of $1/\sqrt{2}$ and searched over all scales. The descriptor matching of the features is done using spill trees which is an alternative to the k-d tree with the Best-Bin-First strategy used in [2]. Here only a single leaf within each tree is visited when matching the descriptors. The resulting candidates are then merged. To estimate the homography the PROSAC scheme is used between the planar target points and the input image. The estimated set of inliers is considered as the final homography which can be used for 3D pose refinement.

The tracking system is similar to the approaches in [1] and [2]. To find the exact feature location the normalized cross correlation (NCC) approach is used. Additionally a motion model is used for each target to calculate the camera pose. It is a linear model which calculates the pose of previous two frames to predict the current pose.

4.4. Analysis of A Learning Based Approach

It is based on bi - layered clustering approach which allows the system to index and retrieve the object based on their tracked features. It is proposed by Pilet J, Saito H. in [4] in which the image retrieval technique is similar to the previous systems described in [1], [2] and [3]. First the features are detected using SIFT or MSER , which generates a set of vector for the images. The vectors are quantized which turns the features into visual words. Once the bag of words are generated the information can be indexed and searched.

The inconsistency of descriptor over the frames is tracked using keypoints from frame to frame. Thus during the training phase the features are captured. The collected data helps generate a stable visual vocabulary. Indexing and retrieval is based on the TF-IDF weighing scheme for the bag of words model. Similar to previously described systems, NCC is used for localization of features from frame to frame and the RANSAC algorithm during the detection process handles the outlier rate. The same process is done by the least median of squares (LMedS) algorithm during the tracking process. During runtime the system matches the point tracked from a single frame with the training data.

4.5. Analysis based on SIFT technique for multiple object tracking

The system proposed by Kiyoung Kim , Vincent Leptit and Woontak Woo in [5] is similar to the approach described for the SIFT technique in [2]. But the approach is based on tracking of multiple objects at the same time using image retrieval and Structure from Motion technique as opposed to [2] where a single object is tracked at a time. The system also allows the user to include new objects for tracking. The tracking process performs frame to frame tracking while the detection process which is slower is used to detect the visible target from the camera. Like the approach described in [2] the SIFT features are extracted from the input camera image which is then run through the vocabulary tree algorithm to retrieve similar images. Kd-tree is used to match the keypoints of the input frame and keyframes. The RANSAC algorithm is used to compute the rotation and translation between the matches. After the object detection the inliers computed from visible objects are send to tracking process. The newly detected points replace the tracked points based on the SSD score. The pose is initially obtained by the P-n-p and RANSAC algorithm and is updated using the Gauss -Newton algorithm. The initial map is generated, based on the PTAM approach.

In the interactive modeling process the objects are tracked based on the keyframes of the objects. The local coordinates of the object are defined by the user. The transformation matrix is computed by placing a 30X30 pixel plane to the keypoints. The user next draws the facet and extends it on to the 3D plane. The system manages two vocabulary tree, one for Structure from motion and another for multiple object tracking.

5. Analysis Of Different Technique For Interactive Modelling

5.1. Analysis of Probabilistic Feature-based On-line Rapid Model Acquisition Technique

The reconstruction technique proposed by Pan Q, Reitmayr G, Drummond T in [6] is designed to simplify and fasten the reconstruction process. It performs online reconstruction of the object held by user using a video- camera. Partial model of the object is generated quickly. ProFORMA can be used in the on-line reconstruction of textured object and for pose tracking. Live video is sent to the tracking module to track object position at frame rate based on the partial model constructed by reconstruction module. The reconstruction module depends on the keyframe for generating the object. The complete model is fed to the tracker to update the model. The visualization module

allows to view the augmented object of the texture model on to the current pose of the object in the live video feed.

To discretise the viewing sphere of an object, an icosahedron is placed at the center of mass. It allows to enable score at each face to represent the uncertainty of how the object looks. If there is high viewpoint uncertainty score the orientation has to be revisited else the orientation has enough information. For every face of the icosahedron the uncertainty score is calculated once for every new keyframe. The direction of rotation of the object can be chosen based on the score computed which is indicated to the user and thus next orientation would be in the direction of the face with the highest uncertainty.

5.2. Analysis of the 3D sketching technique

The system proposed by Simon G in [7] is based on the principles of the Google SketchUp software. The camera acts an interactive device for the system. The modelling module depends on the rotation of the camera. The camera parameters are set before the scene geometry is generated. For camera calibration a pair of horizontal parallel lines are placed orthogonal to each other. New faces are created on contact by snapping the clicked points with the existing 3D points. If new 3D vertices are needed they are generated by inverse ray intersection technique.

The camera pose is estimated using the image content and the modelled planar polygon. The Harris corner detection technique is used to match corners frame to frame using a normalized cross correlation search patch. The RANSAC algorithm is used to reject the outliers. The camera pose is estimated iteratively using the sum of square transfer error (SSTE) which reduces the inlier matches. Tracking failures can be detected automatically by testing if the number of inliers has fallen below a threshold. The pose recovery is done based on SIFT feature matching as it is invariant to image scale and rotation. During the recovery procedure the texture faces are traversed until one face matches the SIFT features extracted from the current frame.

5.3. Analysis of VideoTrace Technique

It is a system proposed by Van den Hengel A, Hill R, Ward B and Dick A. in [8] which is used to generate realistic 3D models of the objects in videos. The shape of the object to be modeled is traced using one or more frames from the video. It helps to generate parts of the screen with the level of information available about the object. The system follows a sketch based interface to create the 3D model of the object through tracing. Initially a 3D point cloud is generated and overlaid on the input video through structure and motion analysis. To generate number of small clusters of adjoining pixels the frames of the video are segmented

The polygon face is modeled as the user traces the boundary within the video frame. The user can also navigate to refine the model. The outline of the 3D model is projected on to a new frame to allow the user to make the necessary adjustments. The 3D position and orientation of each face is estimated by fitting planes to the face which helps reconstruct the 3D points. To ensure robustness multiple planes are fit to the subset of 3D points. The planes are combined to form a hypothesis model of the object. The system includes two modeling modalities based on Nonuniform Rational B-Splines (NURBS). They allow the user to define curved lines over the images. Once the NUMBS curve has been defined the least median of square technique is used to constrain the NUMBS curve to lie on the plane.

6. Comparison Of The Different Methods

An evaluation of the different techniques used for detection and tracking and for interactive modelling is provided below. The evaluation is done based on different factors. Table 1 is an evaluation for the different techniques used by the detection and tracking module. The first column is an evaluation of the SLAM algorithm implementation in PTAM. It is used to create a map of the unknown environment. To create the map the set of keyframes from the video is considered as the input. The output is a map of the 3D point features. The implementation however requires powerful computation hardware and the system has limited live experience. Also, due to motion blur it may decimate most corners features in the image and it has no notion on self occlusion. Another drawback of the approach is that the map is usually generated within a small workspace environment. The k-d tree algorithm is used to match the features of the input frame and the reference frame in the vocabulary tree. The drawback in the approach is that it can track only one object at a time. The approach defined in [3] optimizes the score evaluation between the detection and tracking technique. It can handle large number of objects simultaneously, but is limited to the number of objects visible. The approach defined in [4] is bi-layered based clustering retrieval and tracking method. It can also detect a large number of objects but the system depends on the texture of the object.

Factors	[1]	[2]	[3]	[4]	[5]
Goal	To estimate camera pose within an unknown environment	Track predefined planar objects for a real time system	Estimate pose within low-end devices like mobile phones	Tracking multiple objects and image retrieval	Track multiple objects and perform interactive modelling
Approach	Semi automatic	Automatic	Automatic	Automatic	Semi automatic
Methodology	SLAM algorithm for building the map of an unknown environment . Five point algorithm to initialize the map and RANSAC algorithm to generate the base map	In the pre-processing procedure SIFT features are extracted from the target object and a vocabulary tree is constructed. In the online detection the SIFT features extracted are compared to reference data	A modified SIFT feature tracking is done where the DOG algorithm is replaced by the FAST corner detection	Tracking and image retrieval based on bi - layered clustering method	Based on SIFT feature extraction from kayframe and input frame. Construction of vocabulary tree for the keyframe. Comparison between two frames done using kd tree
Input	Set of keyframes from the input video	SIFT features extracted from the input video frame	Features extracted from the input video	Keypoints tracked frame to frame	SIFT features extracted
Output	Tracked handheld motion and map of 3D point features	Tracked predefined planar object	Tracked planar object	Tracked object augmented with a virtual element	Object tracked based on the user defined target by interactive modeling
Performance	For a cluttered desk 57 keyframes can be generated with 4997 point features	Runs for more than 125 frames per second with 314 planar target	Can track 6 planar target at 23 frames per second	Can track multiple objects with low delay and more than 300 entries in database	Can track 50 different 3D objects in 6 - 35 ms per frame

Table 1 : Analysis of the detection and tracking module

The process of object detection and feature tracking run parallel in [5]. The system maintains the real time performance and can simultaneously track a large number of objects. Table 2 is an evaluation of the different techniques for interactive modelling .The method in [6] can build complex 3D models of the real objects. The AR system is used for reconstruction process but does not track the object later. The approach in [7] combines automated reconstruction and interactive modelling. 3D primitives are fit to the 3D locations of the model. The VideoTrace system described in [8] is an offline reconstruction process with PTAM in [1] for the interactive modelling purpose. The approach in [5] is similar to the approach described in [1] where a second background thread is used to keep track of the camera and the reconstructed 3D location.

Factors	[6]	[7]	[8]	[5]
Goal	Generate 3D model	Acquire 3D geometry from the arbitary scene	Interactively generate realistic 3D model	Interactive modeling and tracking of multiple object
Approach	Semi automatic	Automatic	Semi automatic	Semi automatic
Methodology	ProFORMA system to speed up reconstruction process	Based on the principles of Sketch Google UP and recovery procedure based on SIFT feature extracted	VideoTrace system that allows the user to define the primitives of the object	The 3D primitives are defined by the user to which the planar facet is set and stored in the database as a new target object
Input	Live video sequence	Images ,videos frames etc	Video frame	SIFT features extracted
Output	Realistic 3D model	3D model of the object	3D model of the traced object	Augmented 2D object
Characteristics	Generation of a complete 3D texture model in one minute	The process is tedious and requires some time to develop the model	Easy to create the 3D models in videos	The users are able to create various models for which the tracking data is immediately generated

Table 2 : Analysis of the Interactive modeling technique

7. Conclusion

Based on the literature analysis conducted above, it is found that the system proposed in [5] is the faster and better approach among all the methods analyzed. The system combines techniques of image retrieval and Structure from Motion. The features are extracted using the SIFT algorithm as it is scale invariant. The vocabulary tree is used to match the features between the input frame and the keyframes. The system supports polygon and circular shaped models and can perform multiple object detection. The system can track 50 objects in 3D in 6- 35 ms per frame.

The system also allows the user to add new objects quickly. The region selected by the user in the 2D region is fit with 3D primitives which is adjusted to create the 3D object. The tracking data is immediately generated for the 3D object. The proposed system can be used in a number of AR applications.

8. References

[1] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. In: Proceedings of the international symposium on mixed and augmented reality; 2007. p. 225–34

[2] Kim K, Lepetit V, WooW. Scalable real-time planar targets tracking for digilog books. Vis Computer 2010;26(6–8):1145–54.

[3] Wagner D, Schmalstieg D, Bischof H. Multiple target detection and tracking with guaranteed framerates on mobile phones. In: Proceedings of the international symposium on mixed and augmented reality; 2009. p. 57–64

[4] Pilet J, Saito H. Virtually augmenting hundreds of real pictures: an approach based on learning, retrieval, and tracking. In: Virtual reality; 2010. p. 71–8.

[5] Kim K, Lepetit V, WooW. “Real - time interactive modelling and scalable multiple objet tracking for AR"2012;945-948

[6] Pan Q, Reitmayr G, Drummond T. Interactive model reconstruction with user guidance. In: Proceedings of the international symposium on mixed and augmented reality; 2009. p. 209–10

[7] Simon G. In-situ 3D sketching using a video camera as an interaction and tracking device. In: Eurographics; 2010. p. 53–6

[8] Vanden Hengel A ,Dick A ,Thorm¨ahlenT ,WardB ,TorrPH S .VideoTrace :rapid interactive scene modelling from video. ACM Trans Graph 2007;26(3), http://dx. doi.org/10.1145/1276377.1276485.

[9] Lowe, D., Distinctive image features from scale- invariant keypoints. Int. Journal of Computer Vision, Volume 60, Issue 2, pp. 91-110,2004

[10] Ozuysal, M., Fua, P., Lepetit, V., Fast keypoint recognition in ten lines of code. In of Proc. CVPR 2007, pp. 1-8, 2007