https://doi.org/10.2352/ISSN.2470-1173.2018.13.IPAS-463

Discrimination of active dynamic objects in stereo-based visual SLAM

Ihtisham Ali, Olli Suominen, Atanas Gotchev ; Tampere University of Technology, Tampere, Finland

Abstract

Over the years, the problem of simultaneous localization and mapping have been substantially studied. Effective and robust techniques have been developed for mapping and localizing in an unknown environment in real-time. However, the bulk of the work presumes that the environment under observation is composed of static objects. In this study, we propose an approach aimed at localizing and mapping an environment irrespective of the motion of the objects in the scene. A hard threshold based Iterative Closest Point algorithm is used to compute transformations between point clouds that are obtained from dense stereo matching. The dynamic entities along with system noise are identified and isolated in the form of outliers of the data correspondence step. A confidence metric is defined that helps in identifying and transitioning a 3D point from static to dynamic and vice versa. The results are then verified in a 2D domain with the aid of a modified Gaussian Mixture Model based motion estimation. The dynamic objects are segmented in 3D and 2D domains for any possible analysis and decision making. The results demonstrate that the proposed approach effectively eliminates noise and isolates the dynamic objects during the mapping of the environment.

Introduction

In recent years, the approaches pertaining to Visual Simultaneous Localization and Mapping (SLAM) have been developed significantly; although it is a relatively new field. The research in this field was significantly aided by the release of Microsoft Kinect RGB-D (Red, Green, Blue, and Depth) camera. This field has proved to be of great interest to research and business minds alike, due to its impact applications. The state of the art methods are now capable of running the application in real time with robust performance. However, much improvement needs to be done towards handling problems such as expanded spatial volume with loop closure [1], dense mapping [2], and managing dynamic objects in a scene [3].

A variety of SLAM implementations exist. Each implementation may adopt a different type of sensor or methodology. A typical SLAM approach relies on the Iterative Closest Point (ICP) for registration of point clouds, and loop closing techniques for drift compensation [4]. Apart from RGB-D sensors, simple time-of-flight (TOF), monocular and stereo cameras can also be used for obtaining point clouds. Each of these sensors has its own advantages, coupled with inherent data processing challenges.

Until recently the core assumption for SLAM has been that the environment under observation is static, i.e. none of the observable objects in scene propose any change in their dynamics or shape.

As a result, this assumption leads to inconsistent map, erroneous localization, residual noise and possible failure in registration, when the environment is dynamic. Nevertheless, a few studies have successfully dealt with dynamic objects in the scene. Many of these studies use Kinect to obtain the depth maps [5].

Typically, dynamic objects in a scene can be detected and isolated for SLAM using CAD models or other form of prior knowledge with the use of commercial RGB-D sensors. However, such an approach limits the applications of the system. In this study, we demonstrate the application with a stereo camera for localizing and mapping an active dynamic environment without any prior knowledge about the dynamics in the scene.

Related Work

Davison et al. [6] introduced a real-time camera tracking system known as monoSLAM (monocular Simultaneously Localization and Mapping) to localize and map a freshly explored environment. It uses an extended Kalman filter (EKF) to estimate the camera pose. Later, in [7], Newcombe and Davison adopted structure from motion (SFM) to find the ego-motion and reconstructed a detailed model of the environment. Along with the prior mentioned studies, the approaches presented in [8], [9] and [10] maintain the assumption that the underlying environment is stationary and suggests to discard the dynamic element points by considering them outliers to the systems.

Nonetheless, the research for developing SLAM algorithms in a static environment has matured considerably. Hence, many researchers are now focused on implementing SLAM in a dynamic environment. In [11] Andrade-cetto et al. used a stereo camera to build a map for mobile robot localization. It uses strength augmentation of features and robot localization to learn in a moderately dynamic indoor environment. The landmarks used for mapping are low (approx. 50) and provide little information about the nature of the environment. In an attempt to capture more information about the dynamic object, Aguiar et al. [12] suggested a multi-view camera approach. This technique utilizes eight cameras to track a person and reconstruct a spatio-temporally consistent shape, texture, and motion of the performer at a high quality. Through a different approach, Zollhofer et al. [13] proposed to reconstruct a non-rigid body in real time with a single RGB-D camera. The non-rigid registration of RGB-D data to the template is performed using an extended non-linear As Rigid As Possible (ARAP) framework by implementing on an efficient GPU pipeline. Unfortunately, like many other implementations, [12] and [13] require an initial static model/template of the body that is later tracked and reconstructed. The template is then deformed over time based on the rigid registration and non-rigid fitting of points. However, the limiting factor is that the spatial extent of the scene is limited to a single object of interest. Additionally, the system may fail at registration and tracking in case of occlusion, sparse or noisy data.

The aforementioned limitations were successfully removed by Keller et al. [14]. The authors proposed a Point-Based Fusion approach to reconstruct a dynamic scene in real-time using Kinect/PMD Camboard. The approach considers outliers from ICP as possible dynamic points and assigns a confidence value which later determines if the point is static or dynamic. The dynamic points are used as seeds for region growing method in order to segment the entire dynamic object in its corresponding depth map. The implementation proves to work effectively as it can reconstruct both static and dynamic scenes at a considerably good quality. Unlike the previous methods, it can map a comparatively larger spatial area and has been tested in an indoor environment. However, the use of these commercial RGB-D cameras is only suitable for the indoor environment, and it is very difficult to obtain meaningful data even with Kinect V2 in an outdoor environment. The maximum range of depth camera in Kinect V2 diminishes to 1.9m under the most favourable conditions with only two-thirds of the data being reasonably accurate [15]. In the presence of direct sunlight, the operation range falls below 0.8m [15]. The stated figures were obtained empirically with the data being processed and effectively denoised for better operation [15].

System Overview

An overview of the proposed approach is shown in Figure 1 in the form of a flowchart. The process pipeline takes in stereo images to compute the disparity maps and reconstruct 3D points after preprocessing. A six degree of-freedom (6DoF) pose of camera is computed for consecutive steps using the ICP in order to transform the 3D points from camera coordinates to a global map based in the global coordinate system. The outliers of the data association are not discarded. Instead, the outliers are used to understand the dynamics of the objects in the environment under observation. A Gaussian Mixture Model (GMM) based motion estimation method is used to corroborate the results obtained for the dynamic environment. The input data to GMM is preprocessed to extend its validity to moving sensor applications.

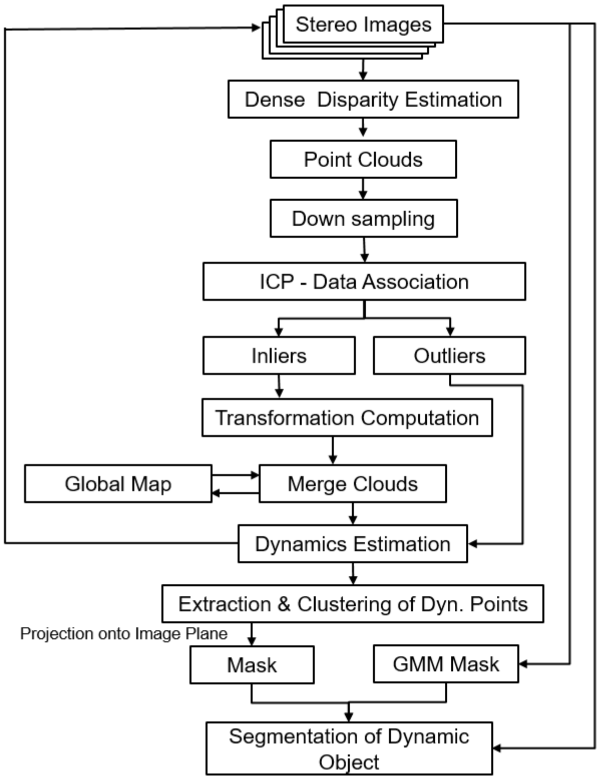

Stereo Images
Dense Disparity Estimation
Point Clouds
Down sampling
ICP - Data Association
Inliers
Outliers
Transformation Computation
Global Map
Merge Clouds
Dynamics Estimation
Extraction & Clustering of Dyn. Points
Projection onto Image Plane
Mask
GMM Mask
Segmentation of Dynamic Object

***Figure 1**. System pipeline*

Approach

Preprocessing

The stereo pair obtained for a scene is used to compute disparity estimates after rectification of the images. We employe the approach of dense disparity computation compared to sparse feature matching. Though feature matching based approaches can provide more consistent and accurate depth estimates, for pose estimation based on ICP and applications like 3D reconstruction, dense disparity estimate prove more useful. The disparity maps were computed using the Semi-Global algorithm as it offers a good compromise between computational speed and global optimality. Each pixel position $(x, y)^T \in R^2$ has its computed disparity $D_i \in R$. The disparity maps are obtained for both the images and a consistency check is performed from one camera to the other in order to remove false disparities. The 3D positions of the valid disparity points are recovered in the form of dense points clouds. However, to ease the computation and memory complexities for SLAM, the point clouds are uniformly downsampled using a grid filter. Each individual point cloud PtC_t has the associated description of each point i.e. Location (X_k, Y_k, Z_k), Color (R_k, G_k, B_k) and Normal vectors to the plane (Nx_k, Ny_k, Nz_k) stored along with it.

Data Association and Pose Estimation

The data association and pose estimation are the constituent steps of point cloud registration. During the registration step, the points from PtC_t are searched for correspondence with points from PtC_{t-1}.The algorithm Iterative Closest Point (ICP) is used to select the optimum points by iteratively minimizing the error metric e^i given in equation (1).

$$e^i = \sum_{i=1}^{N} d_s^2(T^i p_k\ ,\ S_j^k) \qquad (1)$$

where d_s is the signed distance from a point to the plane, T^i is the transformation computed in the iteration i of the error minimization process, p_k are points from PtC_t and S_j depicts the tangent plane of at point q_j for the points in PtC_{t-1}.The transformation matrix T_t depicts the 6DOF camera pose change between the time t and t-1, where T_t is composed of a rotational matrix $R_t \in R^3$ and translational vector $tr_t \in R^3$.The 3D points and the associated normal are converted to global coordinate using the transformation matrix.

Generally, a percentage of closest points are selected as inliers for minimizing the error metric and computation of camera pose. However, we adopted a hard threshold based approach that filters the nearest points selected in each iteration, thereby removing most of the wrong correspondences (outliers) from the process that are present either due to noise (erroneous depth estimation, different sampling of an entity or motion of the objects. The points that help to obtain the correct transformation during the iterative process are known as the inliers.

Once spatially transformed, the new point cloud is merged with the global cloud or the 3D map. The global cloud in our work stores additional two descriptions for each 3D point in addition to the original three properties of a point in a point cloud. A confidence metric C_k and frame presence F_k is defined for each 3D point. The confidence C_k of a point informs us about the integrity of the point for being static and valid while the frame presence F_k stores the information about first time the point was introduced to the system.

Merging and Confidence Gain

The addition of new points in each iteration adds a significant amount of computational complexity and memory load on the processing pipeline. Therefore, it is advisable to remove any redundant or erroneous 3D points from the global map. Merging of points serve as one of the two steps that help in reducing the number of points in the global map. A point q_n in PtC_t may find multiple valid inlier correspondences in PtC_{t-1} during the transformation computation, however, only the closest single point p_n is merged physically after registering of the point clouds. The physical properties (location, color and normal vectors) are averaged to create a new point, which helps to remove the redundant duplicate. The confidences of both p_n and q_n are summed and increased by a constant (0.1 for our experimentation). The frame presence F_k is incremented once for all the inliers, to signify that it has been observed in the scene. In contrast the remaining valid associations (among the inliers) of q_n are not merged physically but added with the original properties to the global map since they might represent a different view of the same object. However, the confidence of these points is merged with q_n and raised through a gaussian distribution as shown in Figure 2. The maximum confidence is set equal to the constant 0.1 for the closest points. The further the points are from each other, the lower confidence it gains. The standard deviation of the gaussian distribution is set to be half of the correspondence threshold used during ICP.

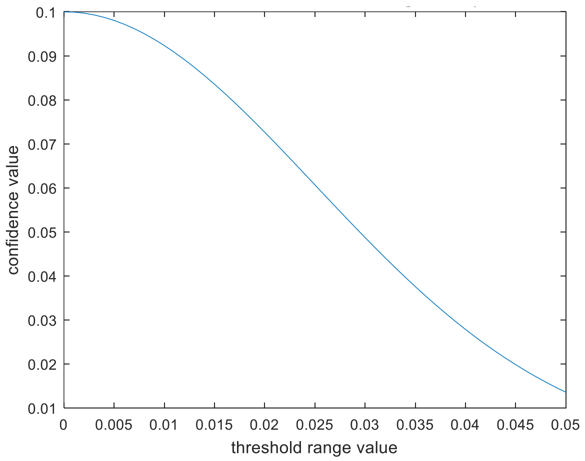


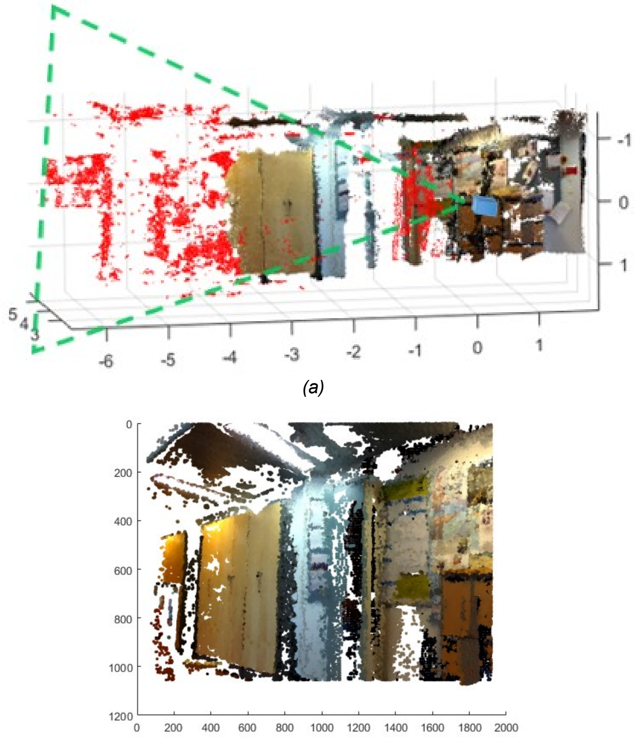
***Figure 2**. Gaussian distribution based confidence assignment at thresh=0.05*

Confidence Reduction and Removal

Confidence gain during merging of the points helped to determine the stability of points in the map. The higher the confidence, the more stable and static the point has been in the scene. However, the reverse is equally important in order to accurately update the global map. If a stationary object in the scene starts to move, the associated 3D points should logically change its position in the global map. The dynamic nature of the points is obtained by continuously reducing the confidence of all the points by a constant (1/10 of the confidence gain in this study) that are in view of the camera and therefore expected to be seen.

The 3D points from the global map are projected to the image plane using the camera intrinsics K and the inverse of global camera pose $T_{g\,t}^{-1}$ at time t. The points that are projected within the bounds of the plane are assumed to be in the camera perspective and, therefore, reduced in confidence. Among these points, those that have been associated with other points would still have a positive confidence change, however, the points that did not find any association would only be reduced in confidence. If the confidence of a point falls below 1, it is assumed to be unstable or dynamic.

In order to maintain and update the map, unstable points representing noise and dynamic 3D points are removed in each timestamp. For this study, the maximum confidence that a point can acquire is 1.25 which was selected empirically while keeping into consideration the confidence gain and reduction values. The global map is searched for these unstable points that have remained unstable for more than a threshold time t_{max} and are removed from the map.



(a)

(b)

***Figure 3**. Projection of points to image plane for confidence reduction (a) illustration of the camera viewing the global cloud (b) projected points onto the image plane from the perspective view*

Dynamic Estimation

The Global Cloud is composed of both static and unstable/dynamic points. The static points have high confidence measure that is obtained through the continuous merging of points from close timestamps. The unstable noise or points from a dynamic object are observe at different position and with less consistency, hence, they do not accumulate enough confidence. It is essential to discriminate the unstable points due to point cloud reconstruction inaccuracy or slightly off localization and the points pertaining to true dynamic entities.

For an image frame at time t, the low confidence points from the global point cloud are projected on to the image using the accumulated transformations $T_{g\,t}^{-1}$ computed during the registration. The points that lie within the bounds of the image plane are indexed and clustered in 3D space based on their distances. For each frame at time *t* multiple *k* clusters $C_{t\,k}$ might be created. Clusters with fewer points than a threshold are removed. This threshold may vary depending on the downsampling of the original point clouds. A 2D mask is generated from the boundaries of the projected clusters on the image plane. This mask may contain the bounds of both dynamic points and the unstable noise. In order to discriminate between the two, another verification step is included in the process.

A GMM based motion detection approach [16] is adopted to highlight moving objects in the scene. GMM is a background modelling technique that is trained on images to learn the background model at pixel level and describe each pixel with K gaussian distributions. The approach detects any moving object that does not fit the model's description in the form of foreground. However, a limitation exists to the direct application of GMM. The approach is only applicable for static camera systems. In this study, we extended the use of GMM with specific preprocessing steps. With a moving camera, the background changes frequently, therefore, the model is continuously trained with few images as function of displacement. For this study we used 3 to 10 images based on the displacement from current scene. Moreover, the training images are geometrically transformed to the current frame by tracking salient features in the images in order to maintain the assumption of static camera for GMM application. The approach provides a clean highlight of the moving objects for small translation between consecutive images. The mask obtained using motion detection with GMM is used as seed to select the blobs from the first mask obtained by the projection of the clustered 3D points. The verified clusters are used for segmentation of the moving object from the images as shown in Figure 4.

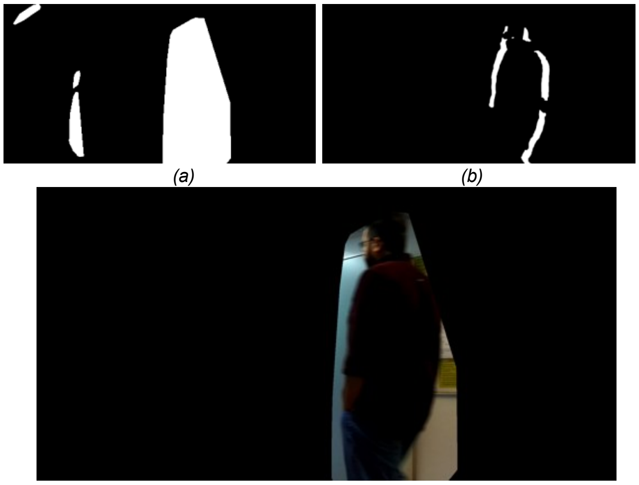


***Figure 4**. Segmentation of the dynamic object using binary masks (a) Mask obtained by projecting the clustered 3D dynamic points onto image plane (b) Masked obtained using GMM based motion detection(c) Segmented dynamic object from 2D images*

Experimental Results

This section provides an overview of the experimentation setup adopted for this study and analysis the results obtained using the proposed approach.

In this study, the data was recorded using a commercial stereo camera Zed [17]. The stereo camera follows the Pinhole camera model with a baseline of 12cm between the camera pair. The standard specifications of the camera quote to work both indoor and outdoor with an effective range of 0.5 to 20 meters [17]. However, the accuracy of depth estimation decreases with distance, therefore, we limited our interest to 6.5–7.5 meters during outdoor usage for more consistent and accurate data. The Zed camera was calibrated using the calibration approach provided by Computer Vision System Toolbox™, which is based on the work of Jean-Yves Bouguet [18].

The system is tested on various test scenes of varying dynamic nature to better comprehend the performance of the approach. The scene shown in Figure 5 demonstrates the ability of the system to incorporate dynamic objects in the environment. The scene was recorded at 30fps with the camera being fixed in the environment. The 1st Column of images show the excerpt from the videos sequence; the seconds column shows the updated map/ global cloud and the last column shows the objects segmented objects when in dynamic state. The dynamic points from the moving object are successfully incorporated as part of the map and then effectively updated during the motion.

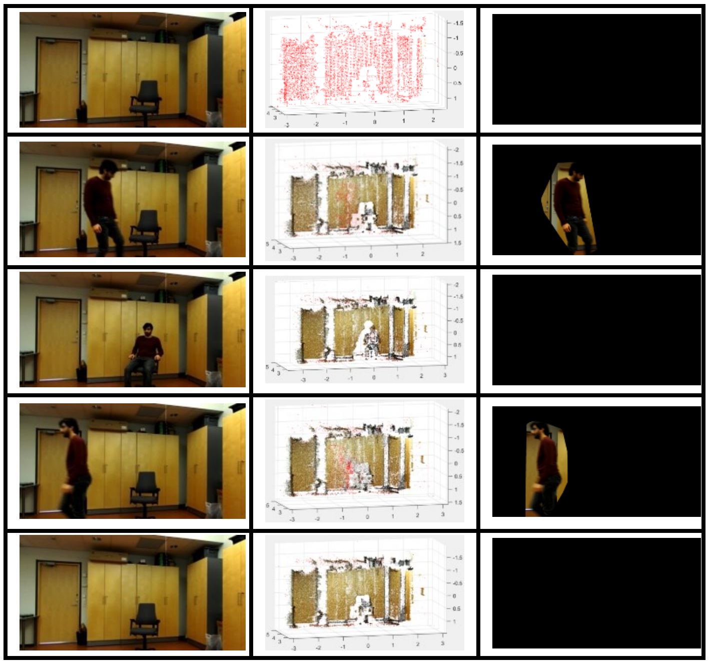


***Figure 5**. Test sequence with stationary camera*

The second indoor scene shown in Figure 6 records a dynamic environment where camera motion is introduced as an additional challenge. The test sequence updates the map while the person passes by in the corridor. The segmentation step not only accurately segments the dynamic object in the middle of the scene but also at the far end of the corridor where most of points are unreliable.

The test sequence shown in Figure 7 was recorded in an outdoor environment over a longer time. The scene was recorded on a cloudy winter day. In order to test the robustness of the approach, the video was acquired using a hand-held Zed camera at 10 fps, and as a result, the scene includes sudden erratic motion. It can be observed that the moving objects in the scene are highlighted in the global map and effectively removed once they pass from the scene, however, the static objects such as the tree is retained in the map even though they are exposed for approximately the same period.

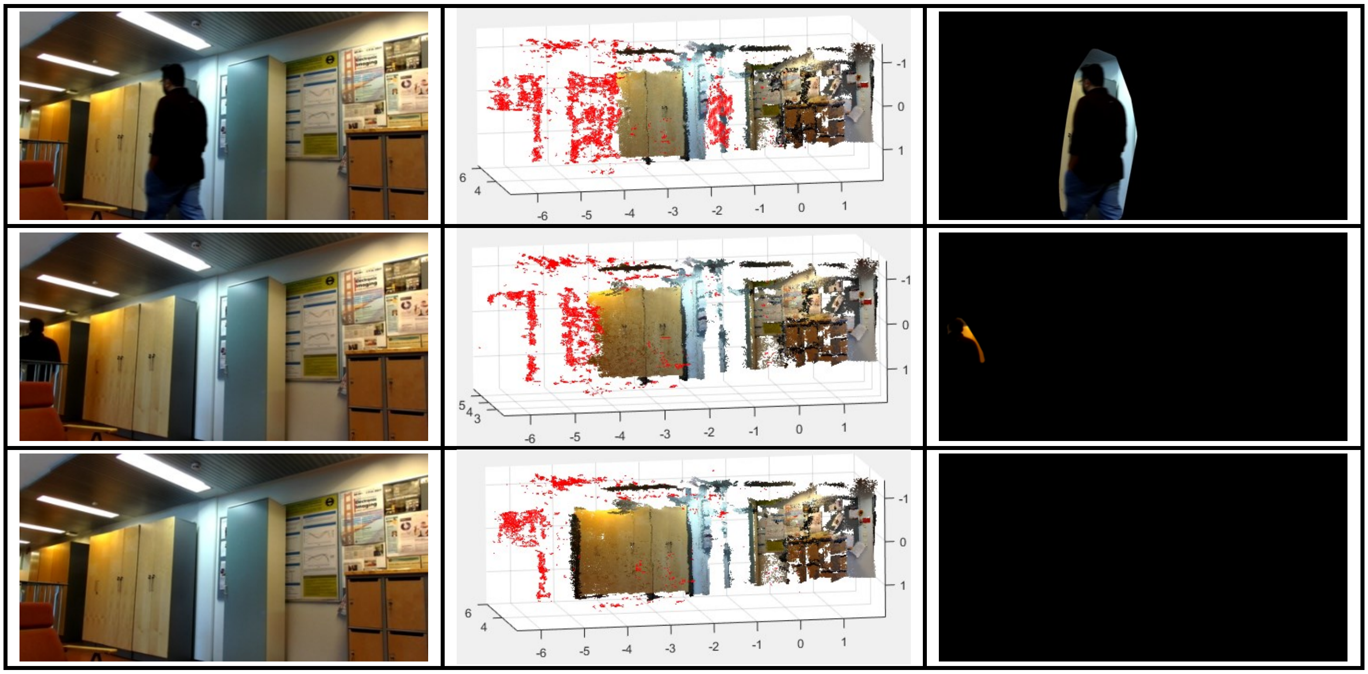


Figure 6. Indoor test sequence with moving camera

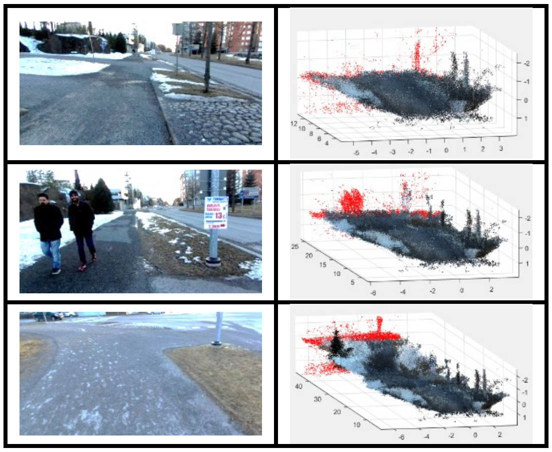


Figure 7. Excerpts from outdoor test sequence with dynamic objects and moving camera

Conclusion

We proposed a scheme to discriminate active dynamic objects present in an environment while localizing and mapping the scene using a stereo camera. The approach is tested on datasets composed of both indoor and outdoor test scenes recorded at various acquisition rates and external challenges such as erratic camera motion, less distinct geometrical structures, and low illumination. The system effectively localizes the observer in the dynamic environment and builds a map irrespective of the relation of motion of camera to the motion of objects in the observed environment. The moving objects are successfully segmented in both the 2D and 3D domains for further extensive analysis.

References

[1] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers and W. Burgard, "An evaluation of the RGB-D SLAM system", in IEEE International Conference on Robotics and Automation, 2012.

[2] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces", in 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007.

[3] K. Litomisky and B. Bhanu, "Removing Moving Objects from Point Cloud Scenes", in Advances in Depth Image Analysis and Applications, Berlin, 2013, Vol. 7854, pp. 50-58.

[4] Bradski, H. Strasdat, J. M. M. Montiel, and Andrew J. Davison. "Scale drift-aware large scale monocular SLAM." Robotics: Science and Systems VI, vol.2, 2010.

[5] Korn M. and Pauli J, “KinFu MOT: KinectFusion with Moving Objects Tracking”, in Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Berlin , 2015, Vol. 3, pp. 648-657.

[6] A. Davison, I. Reid, N. Molton and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1052-1067, 2007.

[7] R. Newcombe and A. Davison, "Live dense reconstruction with a single moving camera", in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.

[8] M. Meilland and A. Comport, "On unifying key-frame and voxel-based dense visual SLAM at large scales", in International Conference on Intelligent Robots and Systems, Tokyo, 2013.

[9] R. Newcombe, S. Lovegrove and A. Davison, "DTAM: Dense tracking and mapping in real-time", in IEEE International Conference on Computer Vision (ICCV), Germany, 2011, pp. 2320-2327.

[10] C. Kerl, J. Stuckler and D. Cremers, "Dense Continuous-Time Tracking and Mapping with Rolling Shutter RGB-D Cameras", in IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2264-2272.

[11] J. Andrade-cetto, and S. Alberto, “Concurrent map building and localization on indoor dynamic environments”, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16, no. 3, pp. 361-374, 2002.

[12] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel and S. Thrun, "Performance capture from sparse multi-view video", ACM Transactions on Graphics, vol. 27, no. 3, p. 1, 2008.

[13] M. Zollhöfer, C. Theobalt, M. Stamminger, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon and C. Loop, "Real-time non-rigid reconstruction using an RGB-D camera", ACM Transactions on Graphics, vol. 33, no. 4, pp. 1-12, 2014.

[14] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich and A. Kolb, "Real-Time 3D Reconstruction in Dynamic Scenes Using Point-Based Fusion", International Conference on 3D Vision, 2013, pp. 1-8.

[15] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter and R. Siegwart, "Kinect V2 for mobile robot navigation: Evaluation and modeling", International Conference on Advanced Robotics (ICAR), 2015.

[16] P. Kaewtrakulpong, R. Bowden, An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection, In Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS01, Video Based Surveillance Systems: Computer Vision and Distributed Processing (September 2001)

[17] ZED Stereo Camera [Internet]. Stereolabs.com. 2017 [cited 10 November, 2017]. Available from: https://www.stereolabs.com/

[18] Bouguet, J. Y. Camera Calibration Toolbox for Matlab. Computational Vision at the California Institute of Technology. Camera Calibration Toolbox for MATLAB.