Chapter 5

3D Scene Reconstruction for Augmented Reality Applications

Summary

An autonomous system needs to build a semantic 3D representation of the world in order to navigate around within the environment and interact with various objects. We have understood so far that it is important for such a system to localize various objects in 3D for a semantic reconstruction of the scene. This chapter will discuss a potential embedded application of an improved 3D object localization achieved by combining the methods discussed in Chapter 3 and Chapter 4. We would discuss the idea of using an Augmented Reality (AR) application based on a hand-held device like a mobile phone that allows users to virtually map the real environment in 3D. This chapter will discuss how the proposed methods in the previous two chapters can be combined to provide more meaningful 3D mapping of the environment by presenting some visual results in this direction.

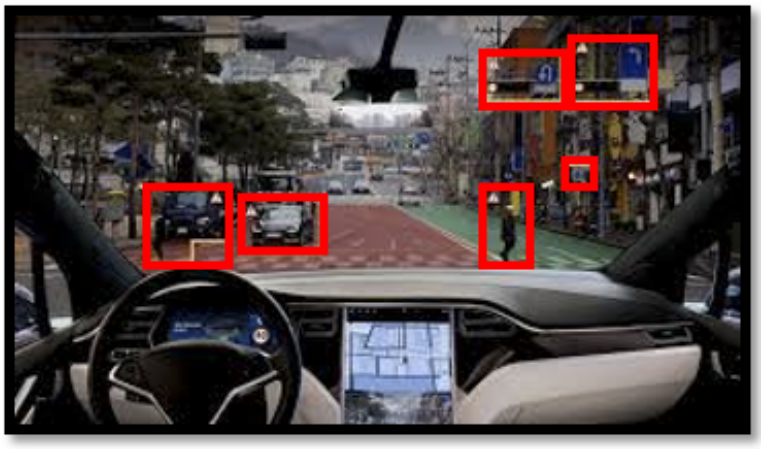

Figure 5.1: An autonomous agent like a self-driving car needs to localize itself and other objects such as people, vehicles and road signs etc. while mapping the environment in which it is navigating at the same time.

Consider a self-driving car navigating on the streets without any human intervention. The car should be able to detect people and other vehicles while at the same time create a map of its own surroundings (e.g. street, road signs, traffic signal, other vehicles etc.) in which it is navigating (Figure 5.1). The problem of building or updating a 3D map of an unknown environment in which an autonomous agent is navigating while simultaneously keeping track of its location within that environment is formally known as Simultaneous Localization and Mapping (SLAM) in the Computer Vision literature.

While a self-driving car is a much more challenging application which involves fusion of data from multiple sensors and performing a lot of different tasks simultaneously, we can think of a personal robot inside a house mapping virtually the entire indoor scene while also navigating itself around. In addition to the 3D mapping and localization, some applications might allow the users to augment the virtual content on top of the real-world scene. An Augmented Reality (AR) platform embedded over a mobile device, for example, would provide the ability to experience the virtual world on top of the live camera image being captured

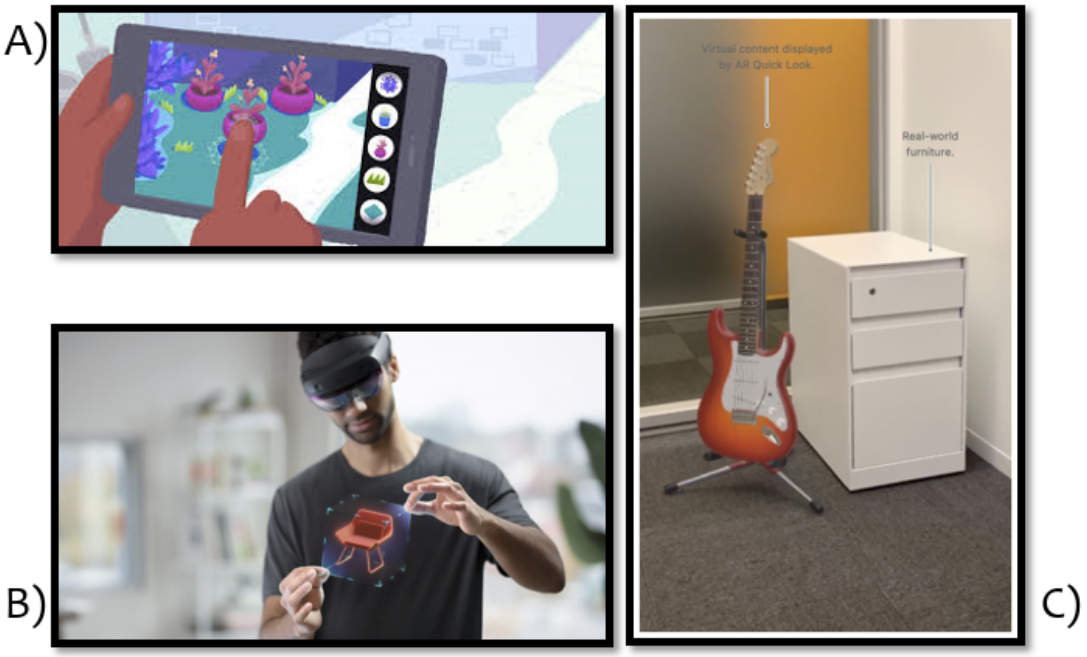

Figure 5.2: A) Google's ARCore [11], B) Microsoft's Hololens [12] and C) Apple's ARKit [13] are some of the popular commercially available embedded software platforms embedded on mobile devices such as phones or tablets for Augmented Reality-related applications.

by the user. We envision building a robust 3D object localization pipeline by combining our proposed methods re-OBJ (Chapter 3) and PSfMO (Chapter 4) within such an AR application which is commercially available and can be easily embedded on any hand-held device like a mobile phone or a tablet.

There are several commercially available AR-enabled mobile platforms such as Microsoft's Hololens [12], Apple's ARKit [13] and Google's ARCore [11] based on its predecessor, Project Tango [130] that allow easily accessible over hand-held devices for the users to virtually map their surroundings, add virtual content inside it and experience it live on their cameras with the illusion of those virtual objects being part of the real world.

For our experiments, we use Google's Tango platform as per the availability, however, any other embedded platform that provides camera pose information

can be used instead of Tango. To allow the users render a virtual world over the real world, the Tango platform needs to be aware of its position relative to the world for which an online odometry and mapping process is used. In this direction, the Tango platform performs two basic functionalities which are mostly common to all AR-enabled mobile platforms:

Motion tracking The motion tracking module uses the camera in the mobile phone device to capture images, find feature points in the images and track them over time to estimate the change in the location of the device. This module also uses an Inertial Measurement Unit (IMU) in addition to the visual information from the camera. An IMU unit consists of accelerometer and gyroscope which gives the position and the orientation of the camera with respect to the world over time. The software then is able to render the 3D virtual world over the real world by aligning the pose of the virtual camera with the pose given by the IMU unit of the device.

Understanding the environment Using the camera of the mobile phone device to detect point features or planes in the images, the Tango module uses this visual information to detect flat surfaces in the scene. Using point features or a cluster of point features, this module tries to detect horizontal planes like the top of the table or the vertical planes like the wall inside a room and also the boundaries of each of these planes to allow users to render any virtual 3D object on the real flat surfaces that appear in the images. However, because of the limitations of the point features, this module might not always be able to detect a plane such as a plain white wall.

As we have pointed out that since the Tango's inbuilt module for understanding the environment only tries to detect feature points in the images to detect planes, it does not map the environment with an object-based representation.

In the next section, we propose a 3D reconstruction pipeline using Tango with

the two methods we have proposed in this thesis for 3D object localization and object instance re-identification already discussed in Chapters 4 and 3, respectively.

5.1 Proposed Method

Using the images obtained from the camera on the Tango device, we can detect objects by employing an off-the-shelf object detector such as YOLO [14]. The objects detected that can be observed across multiple views are used as input to the proposed method, PSfMO for 3D localization.

As shown in Figure 5.3, the PSfMO framework will use these 2D detections observed in multiple views and the camera pose provided by the camera on the Tango-enabled mobile device to estimate the 3D quadrics for the objects' 3D position and occupancy as described in Chapters 4. Since, multiple instances of the same object category might be present in the scene and all of these different object instances are to be tracked across the frames, the re-OBJ system can be utilized to associate all such detections belonging to the multiple instances observed across the views.

We hypothesize that the estimation of 3D quadrics should be improved with the accurate association of all the 2D bounding boxes across the frames. These refined 3D quadrics then can be added to the 3D representation of the scene which can be easily transferred from the 3D point cloud representation shown in the Figure 5.3(E) to a virtual 3D ellipsoid added into the real-world images as observed from the camera on the device.

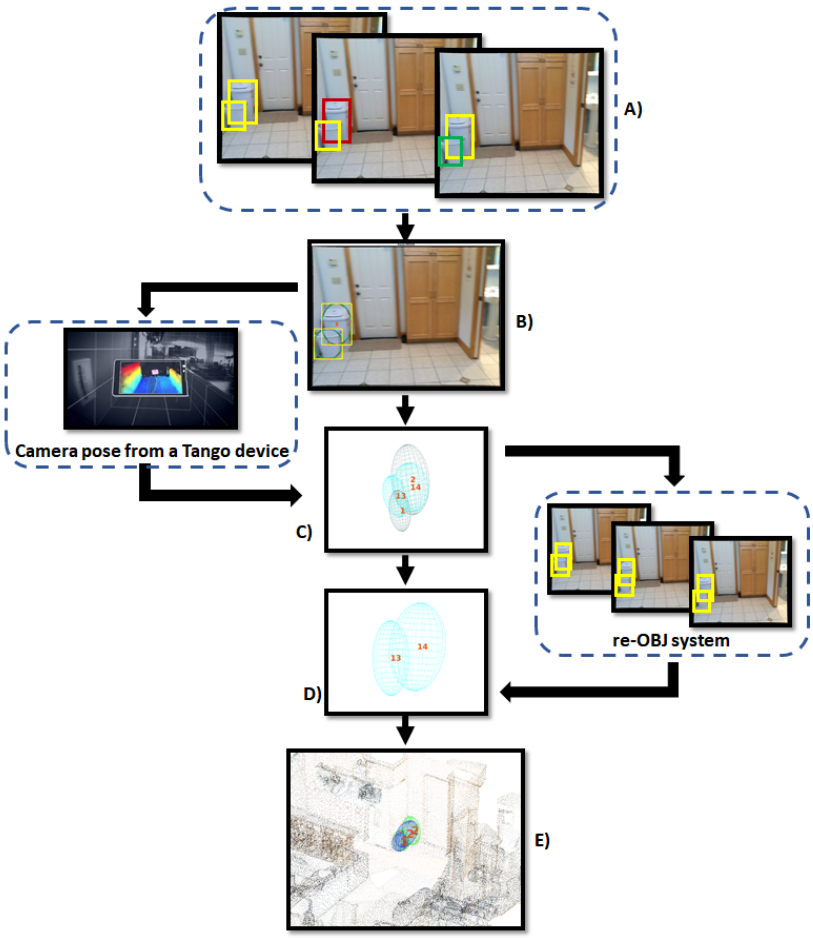

Figure 5.3: A proposed pipeline with a Tango-enabled device integrated with a combination of 3D object localization and object instance re-identification methods presented in this thesis. A) Multiple objects are detected across multiple views. B) These multiple observations can be used as input to the PSfMO framework to estimate the 3D quadrics. C) The camera pose can be provided by the Tango framework for the estimation of quadrics in the perspective settings. If the multiple instances of objects are present and are not correctly associated, it may lead to inaccurate estimation of the quadrics as shown by the grey ellipsoids (The correct ellipsoids are shown in blue color for comparison). D) Using the re-OBJ framework to associate multiple instances of objects across the views, the estimation of the quadrics could be refined. E) The refined estimation of 3D quadrics can be added to the 3D representation of the scene. Here, this representation is shown as a 3D point cloud but in reality, these quadrics can be virtually added to the images captured by the camera on the device.

Figure 5.4: Multiple instances of the same object category, a chair in this case, observed in different views in a video dataset. All these instances are considered to be a different instance at different points in time i.e. different frames in the video.

5.2 Experiments with ScanNet

To understand if the proposed methods are even feasible or not, we present some preliminary results for the 3D quadric estimation on the ScanNet [7] data in case when the object detections are not associated across the multiple views and in the case they are associated correctly in different views.

No object association Consider a static indoor scene where multiple objects are detected across the views in video dataset such as ScanNet [7] as shown in Figure 5.4. All the detected bounding boxes shown in Figure 5.4 are actually the same instance of the object *chair*. However, let's consider a scenario where these instances are considered as different instances of the same object class spread across the frames in the video. This situation is highly likely even when State-of-the-Art methods are used for object matching/tracking in multiple views especially when the camera revisits the same region of the scene at a different point in time.

Figure 5.5: Reconstruction of the scene obtained from the ScanNet dataset [7] mapped with the 3D quadrics representing detected objects in the scene estimated directly by the 2D detections as explained in Chapter 4. Since, from Figure 5.4 we could see that the 2D detections across the images have not been associated as a result of which the 3D quadrics are not estimated accurately (seen on the right in a zoomed version)

Localization from Detection (LfD) method [15] from the 3D localization from detection frameworks such as SfMO [6] and the proposed work in Chapter 4, PSfMO was employed to use these 2D detections to estimate the 3D quadrics.

As can be seen from the Figure 5.5, because of the lack of correct object association across the frames, the estimated 3D quadrics were inaccurate both in the terms of representing the 3D position and also the occupancy of the objects. Now, we will observe the improvement in the 3D localization when the object instances are correctly identified across the multiple views.

With improved object association Now, consider the same static indoor scene with multiple objects detected across the views obtained from the ScanNet [7] dataset.

Figure 5.6: Multiple instances of the object class *chair* observed and correctly identified in different views. All these instances are now re-identified to be the same instance at different points in time.

As shown in Figure 5.6, we consider a scenario where every instance of the object class *chair* is considered to be correctly re-identified across all the views it was observed in.

Assuming the object instance re-identification system proposed in Chapter 3 could correctly associate all the instances of the same object across all the views, we use LfD method again to reconstruct 3D quadrics using the associated 2D detections. As can be seen from Figure 5.7, when the object instances are correctly associated in multiple views, the reconstruction of the 3D ellipsoids became more accurate.

In Figure 5.5, the estimation was poor since the rich geometric information for each instance was lost due to the inaccurate association of the 2D detections. With this poor association, the SfMO algorithm could not obtain much information about the objects in different view points i.e. with a wide camera baseline. On the other hand, in Figure 5.7, the estimation of 3D ellipsoid is more accurate because the SfMO algorithm now has more geometric information incorporated

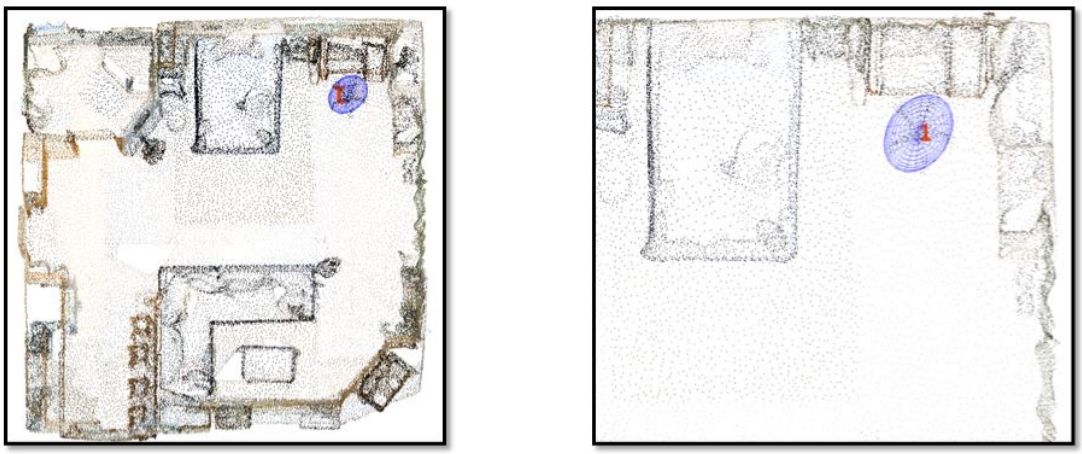

Figure 5.7: We could see from the Figure 5.6 that the 2D detections across the images have been correctly associated due to which the 3D quadric is estimated accurately (seen on the right in a zoomed version) as compared to the 3D quadrics generated in Figure 5.5.

due to the inclusion of many different viewpoints owing again to the correct object association.

Thus, in this section, we found that the object instance re-identification can improve the accuracy of the 3D quadrics estimation method proposed in Chapter 3. The preliminary results on the ScanNet images show promising observations that the 3D localization method such as our proposed method PSfMO and the object instance re-identification can be successfully integrated within the proposed pipeline discussed in Section 5.1 for a semantic 3D reconstruction system using Tango.

5.3 Experiments with Tango data

We used the Tango-enabled android mobile phone, Lenovo Phab 2 Pro [131] to collect the data consisting of images and camera poses. For the purpose of illustration, consider a sequence of images captured from the device as shown in

Figure 5.8: Using the Tango-enabled device, a sequence of images were captured with multiple objects and their multiple instances visible in multi-views.

Figure 5.9: The objects are detected in multiple views using an off-the-shelf object detector such as YOLO [14]. The bounding boxes are fitted with 2D ellipses for the estimation of 3D quadrics as explained in Chapter 4. The bounding boxes are then tracked and associated across the multiple views. This example shows how similar appearances of objects of different categories or multiple instances of the same object category might lead to poor data association.

Figure 5.8. As illustrated in Figure 5.3, the first step in the proposed pipeline is to detect the objects. We used an off-the-shelf, state-of-the-art object detector, YOLO [14] for detecting objects in all the images.

2D ellipses are then fitted within all of the bounding boxes since there is a direct solution from 2D ellipses to the 3D ellipsoids which would represent the objects' 3D location and occupancy in the world. Using the camera pose information from these multiple views, 3D quadrics are generated using the LfD [15] method. A very important step in the LfD method is to track and associate each of the bounding boxes across the multiple views.

However, as can be seen from Figure 5.9, a rigid indoor scene might have many similar looking objects, i.e. objects belonging to different categories but having similar appearance or multiple instances of the same object category. For example, this particular set of images contain multiple instances of *tvmonitor*.

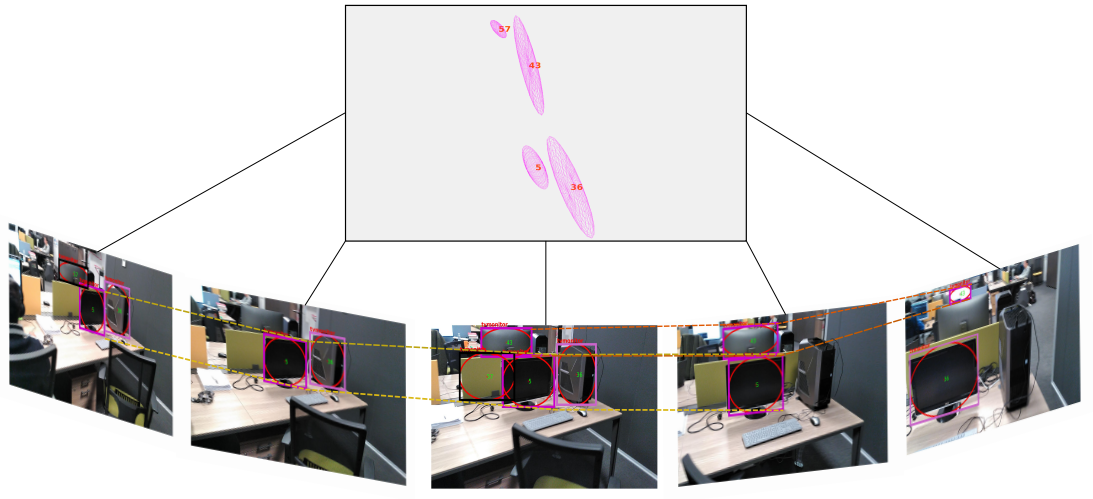

Figure 5.10: Using the camera pose from the multiple views where the objects are detected and an ellipse is fitted within the bounding boxes, 3D quadrics are generated using the LfD [15] method.

Because of the presence of such similar looking instances of objects and some possible false alarms (such as the object ID 57 in the image at the center in Figure 5.9), the association of bounding boxes across multiple views becomes challenging. As we can see, even when a state-of-the-art, multiple object tracker such as deepSort [42] is used for the association, the object Ids from a bounding box at one location in an image switch to another in a different image. Using the 2D ellipses fitted in the bounding boxes in multiple views and the camera pose information obtained from the device, we estimate 3D quadrics (i.e. ellipsoids) as shown in Figure 5.10. Since the association of the 2D bounding boxes was poor, the estimated ellipsoids are mostly inaccurate. As the solution, we used our proposed method re-OBJ to re-identify and match object instances with similar appearance across the frames.

As explained in Chapter 3, the re-OBJ method not only learns the appearance of the foreground but also uses the background around an object instance to learn joint discriminative features that would distinguish this instance with an-

Figure 5.11: Using the re-OBJ method, we discarded the false alarms and associated the 2D detections correctly in multiple views. Here, bounding boxes for two objects are shown for illustration.

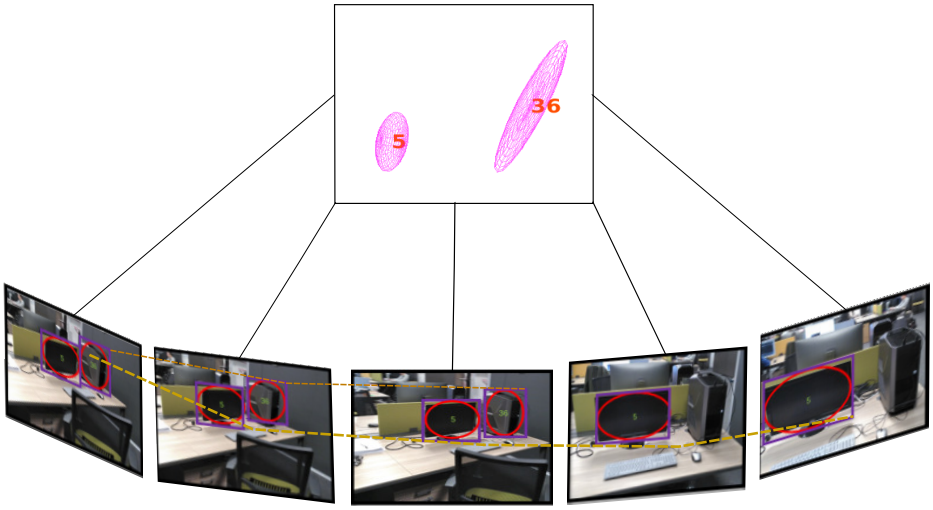

Figure 5.12: Using the camera pose from the multiple views with the 2D ellipses fitted within the better associated bounding boxes, 3D quadrics are generated using the LfD [15] method.

other instance of the same category present in the images. First, we correct the association and then, we use the LfD method again to estimate the 3D quadrics. Thus, from Figure 5.11, we can see the corrected association in multiple views. The 3D quadrics generated after the improved association can be seen in Figure 5.12. In all the frames, the objects shown were correctly associated and thus, the quadrics generated were more accurate than the time when the association of 2D bounding boxes was poor.

The quality of the estimated 3D quadrics can be further validated by back-

Figure 5.13: Once the object are detected, the bounding boxes are associated correctly across the multiple views we estimated the ellipsoids as shown in Figure 5.12. The ellipsoids are then back-projected on the images. The red ellipses in this figure are the ones fitted inside the bounding boxes for the detected object. The yellow ellipse are the projection of 3D ellipsoids generated with the poor association shown in Figure 5.10. The ellipses shown in green color are the projection of the 3D ellipsoids generated after the corrected association of 2D bounding boxes in the images achieved by using re-OBJ method.

projecting the 3D ellipsoids onto the 2D images.

Using the Equation 4.4, we can project the 3D quadrics back onto the images using the camera matrix as can be seen from Figure 5.13.

The figure illustrates how the corrected bounding boxes lead to a better estimation of 3D quadrics which can be clearly understood by the projection of 3D ellipsoids on the images. In Figure 5.13, the 2D ellipses highlighted in red color are fitted inside the bounding boxes. The ellipses highlighted in yellow and green color represent the back-projection of 3D ellipsoids generated with poor association and the corrected association, respectively. We can clearly see that the green ellipses have a better overlap over the red ellipses fitted inside the bounding boxes.

Hence, in this chapter we have seen an embedded application using a mobile device such as Google's Tango platform with an improved semantic 3D scene reconstruction by combining the two methods proposed in this thesis namely, PSfMO for the 3D object localization using only the 2D bounding boxes and re-OBJ, the method to re-identify multiple instances of various objects in the scene.

Chapter 6

Conclusions and Future Work

Summary and Conclusions

In this thesis, we started with a discussion to understand the problem of semantic scene understanding and its various components that help an autonomous agent to build such a system. We discussed how vital it is for any autonomous system to create a 3D reconstruction of the scene.

In this thesis, we reviewed the State-of-the-Art methods in the literature that have proposed different approaches to build a 3D model of the scene in terms of a point cloud by using features like points, edges, curves or planes detected in the 2D images in a Structure-from-Motion (SfM) framework. We found that although the SfM technique has been very successful in modeling the geometric structure of a scene, the point cloud representation lacks any semantic information. An autonomous system that would operate in an environment shared by humans would need to understand the context as well in addition to the geometry of the scene. Thus, in Chapter 1 and 2, we discussed various methods that have been proposed to build an object-based 3D representation of the scene.

We also discussed that the autonomous system needs to localize objects in 3D for adding the required semantic information to the geometric 3D representation of the world. Since there are several methods proposed in the literature that work

directly on the 3D data which is expensive and not so easily accessible, we presented our published work in Chapter 4 that works with 2D data. We presented a probabilistic factorization method to estimate 3D quadrics obtained from 2D bounding box detections observed in multiple views. The 3D quadrics are used as a coarse representation for the objects' approximate 3D location and the occupancy within the 3D reconstruction inspired by the previous work SfMO [6]. We demonstrated that in the case of real-world scenarios, our probabilistic framework (PSfMO) provided more accurate and reliable estimation of the 3D quadrics especially, the occupancy. When the number of views where the objects are detected are less or the camera baseline across the views is not wide enough to capture a great amount of geometric information from different viewpoints, it becomes more challenging to estimate the occupancy accurately using the SfMO method. The main contribution of our method is its ability to use simple statistics collected from 3D CAD datasets to achieve higher robustness to noise and be able to provide improved results even if very few views are available.

Much like the standard SfM methods where the 2D point features detected in multiple views need to be matched, the 2D detections also need to be associated in the SfMO and our proposed framework, PSfMO for the 3D localization task. In Chapter 2, we discussed the various challenges in matching the 2D detections in multiple views and in Chapter 3, we defined a specific problem within the context of static indoor environments where the scene is cluttered with many objects which might have multiple instances present across many different views. We provided a new framework (re-OBJ) for object instance re-identification that jointly learns the foreground and the background of an object instance. The main contribution of this work is that we explored our intuition that the information obtained from the background surrounding the detected target objects in a rigid scene could be highly useful in discriminating two near-identical objects or two

instances of the same object class. Our experiments have shown that the proposed method performs well even in the case of highly cluttered rigid environments like the real-world indoor scenes. Finally, we observed that both the proposed method in this thesis, PSfMO and re-OBJ can be successfully integrated in a commercially-developed Augmented Reality (AR)-based embedded application. The preliminary results show that it is feasible to include the re-OBJ framework along with the PSfMO method for improving the estimation of 3D quadrics and consequently, the 3D reconstruction of a scene.

Potential Future Directions

An open problem in our PSfMO framework is that the estimation of accurate occupancy represented by the 3D quadrics can only be achieved for those objects for which 3D object shape priors would be available for the case where the number of views are less or the camera baseline is not wide enough. As a future work, we propose different kind of prior information that could be tested. First, a classifier with coarse pose information would provide constraint on the ellipsoid orientation. Then higher-level semantic information using scene context [132] and the relation between objects [133] could be extracted to guide the ellipsoid centre estimation.

We also acknowledge an open problem in our proposed work for object instance re-identification. Since, we proposed a joint learning of the foreground and the background to identify multiple instances in the indoor scenarios, we rely on the fact that in video datasets, the background doesn't change drastically with the camera motion. However, it is highly likely that with a wide camera baseline the background around the object might completely change. In such scenarios, we propose to use other objects in the neighbourhood or the relation between the objects [133] for a robust object instance re-identification system.