Dense semantic mapping through deep learning

Wajahat Akhtar, Albert Clrigues and Yu Liu

***Abstract*— Place recognition is an essential component for the creation and use of maps built by SLAM algorithms. In this project, we explore a novel approach for place recognition that uses dense semantic labels as features. Techniques for segmentation of dense semantic labels have only recently achieved acceptable accuracy through the use of Deep Learning. First, a global descriptor based on a spatially aware Bag of Semantic Words is proposed. Then, an adequate matching procedure is suggested that takes into consideration the specific properties of this descriptor. The proposed method is shown to correctly associate places under varied lighting conditions.**

I. INTRODUCTION

Mapping is an essential tool for mobile robots in order to autonomously and safely move through their environment while carrying out their tasks. Simultaneous Location And Mapping (SLAM) algorithms are widely used by robots to create and use maps of their environment. The potential utility of SLAM algorithms not only depends on the accuracy of the constructed map and localization, but also on the kind of information they capture and how convenient its usage is for the development of other higher-level tasks.

Traditional maps with sparse points offer insufficient information for a robot to safely navigate and are difficult for humans to understand. Denser maps can provide rich texture and full structures of environments. However, the large amount of raw information contained make them inadequate for high-level robot intelligence. This requires semantic scene understanding of the surroundings and objects to make the right decisions and take actions. For example, if a robot understands a specific part of the map is a door, a series of actions can be carried out on the door, i.e. opening the door to search for something at the other side. In general, maps built for intelligent robots need to encapsulate geometry, appearance and semantics. Dense semantic mapping has huge advantages for intelligent robots performing higher-level tasks and navigation. The addition of this new feature enables dense metric maps to contain already processed semantic understanding of the scene, which can be directly utilized for higher-level tasks such as safe navigation.

The availability of these additional features incentivices its use for other lower-level tasks that can benefit from its properties. In this project we explore the use of semantic information for a place recognition algorithm with the aim of improving loop closure detection and relocalization for SLAM algorithms. The use of such information for place recognition is considered intuitive and reasonable, as humans seem to rely on a mixture of object-level abstraction and appearance of the environment to localize themselves in the world.

A. Dense Semantic Mapping for map improvement

Semantic information is intrinsically appearance invariant and features derived from it should retain this property. A global scene descriptor based on semantic information could potentially achieve absolute viewpoint and lighting invariance, assuming optimal segmentation.

Dense semantic labels (also referred as DSL in the rest of the report) are obtained by segmenting an image, producing label probabilities for each of the pixels (i.e. wall, door, chair...). Dense semantic mapping embeds DSL to each of the elements of dense metric maps, being widespread the use of point/surfel clouds. To effectively use this feature to improve the quality of the metric and semantic maps created by SLAM algorithms a good understanding of their properties is required. A way to mutually help improve the quality of metric and semantic maps needs understanding of the weak and strong points of both.

The state of the art SLAM algorithms produce metric maps that are locally accurate and consistent and, under partial viewpoint and lighting changes, also globally accurate. The large-scale accuracy of these maps is strongly dependent on loop closure detection algorithms that can robustly add constraints for global map optimization.

Regarding semantic mapping, current state of the art Deep Neural Networks for indoor scene segmentation like SegNet [1] achieve average accuracies of around 50%. Our testing has shown that, although point-wise segmentation accuracy is not very high, the mis-labellings are consistent between several segmentations and generally consistent when looking at the more general object level segmentation.

Consequently, the work developed for this project proposes an algorithm that uses the globally accurate semantic map to perform lighting and viewpoint invariant place recognition, which in turn can be used for relocalization and loop closure detection.

II. RELATED WORK

The reviewed literature on loop closure and place recognition algorithms for SLAM systems is mainly comprised of appearance-based methods.

The Bag of Words approach is widely popular given its flexibility and scalability for outdoor SLAM. The latest developments have been mainly centered around improving the perceptual aliasing and accuracy of Bag of Words under different lighting conditions. The Bag of Word Pairs method proposed by Kejriwal et al. [2] extends the BoW approach to reduce perceptual aliasing by using the relative spatial co-ocurrence of words in the scene. This adds spatial awareness into the BoW approach and has been shown to improve the

recall performance. FAB-MAP [3] uses a Bag of Words approach using SIFT descriptors, which is later complemented by a 3D graph matching procedure that greatly improves the accuracy while reducing perceptual aliasing. However, being SIFT the underlying descriptor it fails under high viewpoint changes or non-affine lighting changes. Extensions to the FAB-MAP algorithm like [4] that propose a hybrid system combining features of FAB-MAP and RatSLAM achieve better lighting invariance to different seasons and times of the day.

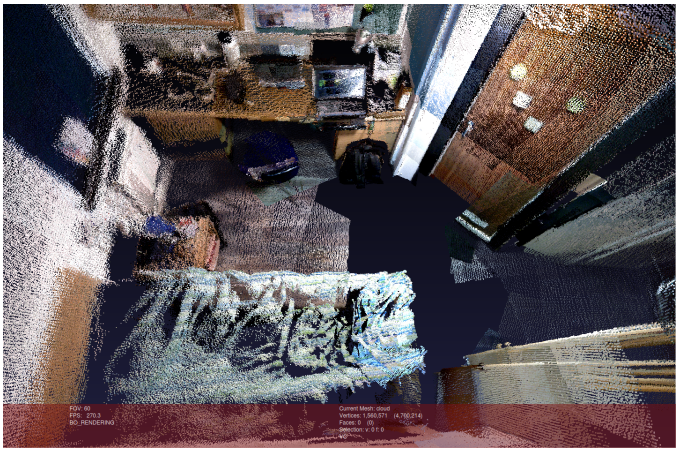
There exist recent approaches based exclusively on Deep Learning [5] that propose CNN based image descriptors for loop closure detection. These descriptors achieve state of the art results without significant light change and outperform the state of the art with significant light changes. The work of Cascianelli et al. [6] features CNN based local features and uses covisibility graphs for the matching procedure to achieve grater viewpoint invariance.

Even though purely appearance-based methods deliver good results and matching accuracy, they still have some intrinsic challenges that compromise their future usability. Firstly, there is still no appearance-based descriptor that achieves large viewpoint change invariance. This means that for robust place recognition a database should contain several images of the scenes under different viewpoints, exponentially increasing the number of elements in the database. Secondly, appearance-based methods, lacking semantic interpretation, cannot differentiate between static or dynamic objects in the scene. As a consequence, the frame descriptors will also include features from dynamic objects that will reduce the recall performance and hence their robustness. Moreover, even if a changing scene is correctly matched, there is no elegant solution to ignore the dynamic parts and the storage of both frames means an even bigger database.

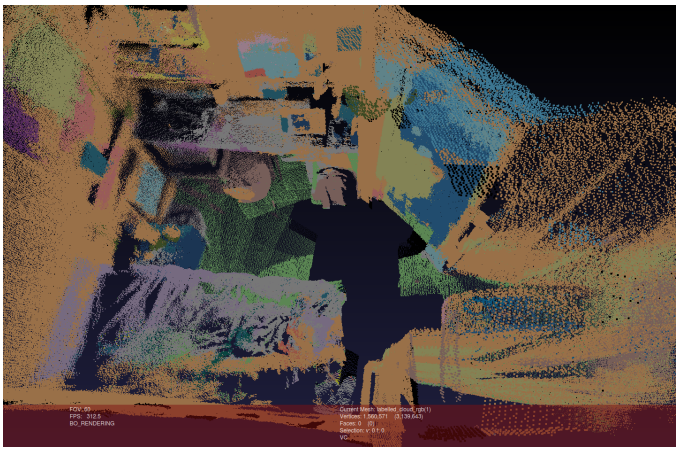
Regardless of the performance of appearance-based methods, there exists an intrinsic benefit on the usage of dense semantic maps for other lower-level tasks. If dense semantic mapping becomes a requirement for intelligent robots to perform their tasks, reusing this same features to carry out other tasks would decrease storage and computational requirements. Greater efficiency can be achieved if it can be avoided that the different algorithms onboard of a robot each keep a different database of features and process information separately.

III. PROPOSED ALGORITHIM

In this section we present our place recognition algorithm utilizing DSL as features. The objective is to compare different places represented by our proposed descriptors of a scene, which include both the frequency of individual semantic labels - Bag of Semantic Words (BoSW) and relative spatial co-occurring semantic labels - Bag of Semantic Word Pairs (BoSWP). The proposed descriptors are extracted from a globally registered 3-D scene (see figure 1b), which is resulted from fusing multiple DSL information segmented from all 2-D key frames following the camera trajectory. A partial framework of fusing 2-D DSL into the registered 3-D pointcloud as proposed by McCormac et al. [7] is applied. We then perform descriptor matching between different places based on computed tf-idf scores and cosine distance. Our place recognition algorithm utilizing DSL features and semantic spatial co-occurrence describes scenes at the object-level. The resulted descriptors suffice to consistently represent different places under varied lighting conditions. This improves from other state of the art appearance-based methods which are vulnerable to non-affine lighting changes.



(a) RGB pointcloud



(b) Semantically labeled pointcloud

Fig. 1: Globally registered 3-D map with DSL features

A. Place Description

Two proposed place descriptors (BoSW and BoSWP) are extracted from the globally registered pointcloud fused with DSL. The BoSW accounts for the frequency describing the occurrence of each semantic class in the scene. In particular, all 37 classes of SegNet [1] are considered except for wall, ceiling and floor as indoor scenes are usually accompanied by these three classes, which do not provide discriminative features. As a result, BoSW is a length 34 $(37-3)$ histogram, where each bin corresponds to the number of point in the scene categorized by a semantic class.

The BoSW measures the statistics of DSL features of the global pointcloud. It is fast, easy to implement and quantitatively provides a fine object-level global description of the scene. However, BoSW easily fails to differentiate between any two places having similar content in terms of object type and quantity the common problem of perceptual aliasing (i.e., two identically furnished offices will have the same BoSW even though having distinct setup).

We propose to resolve this problem by incorporating the spatial adjacency information into our descriptor. This idea was inspired by the creation of Bag-of-Word Pairs as proposed by Kejriwal et al. [2], where they formed word pairs by exploiting the relative spatial co-occurrence of the words. In brief, their words correspond to 128-dimensional SURF descriptors defined by scale size. A word is said to form a pair with another word if the first one's center lies within the spatial neighborhood of the second one, as illustrated in figure 2. In our algorithm, such pairing scheme is modified to adapt to DSL features instead of SURF features.

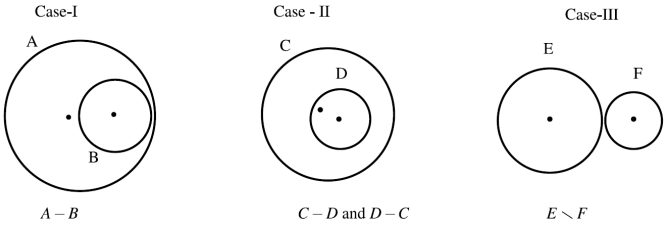


Fig. 2: Understanding relative spatial co-occurrence of words in a pair. While Case I and Case II show a valid word pair, case III does not.

We perform clustering on the global, semantically labeled pointcloud. Clustering is first carried out by semantic class in order to divide points of different classes, and then by point-to-point distance to separate points of the same class that are spatially apart. The resulted clusters having insufficient number of points are considered noise hence not taken into account in subsequent steps. Each remaining cluster corresponds to an object with two properties: its semantic label and a geometric center (i.e., 3-D coordinates with respect to the global maps reference). We then define a bounding sphere around each clusters geometric center to enclose all points of that cluster (i.e., the largest dimension of an object is used as the diameter of the bounding sphere). Object-level spatial co-occurrence happens when the geometric center of one object resides inside the bounding sphere of another one (see Figure 3). As a total of 34 semantic classes are considered (excluding ceiling, wall and floor) and any object can form a pair with any other object (from the same or different class), the BoSWP is a length 1156 (34^2) histogram, where each bin corresponds to the number of pairs in the scene categorized by two semantic pairs. As pairs are defined relative to the spatial location of each cluster center, BoSWP exhibits a directional attribute.

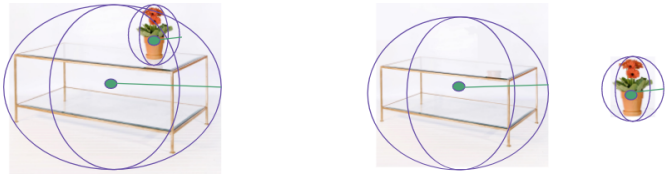


Fig. 3: Understanding relative spatial co-occurrence of words in a pair. Each object's bounding sphere is defined as the largest dimension of the object. Pairs are formed in the left image where the geometric center of one object lies within the sphere of the other one. The right image does not form a pair.

B. Descriptor Matching

The BoSW and BoSWP are two separate DSL-based descriptors (essentially two frequency histograms) to represent a place. Inspired by the matching approach used by Kejriwal et al. [2] and one of their main reference from Angeli et al. [8], we computed a similarity measure between a place and every other places based on the tf-idf (term frequencyinverted document frequency) score using these two descriptors. Two separate observation vectors are computed for each place as shown in equation 1 and 2:

$$z_k^{SW} = \{z_i^{SW}\}, i = 1, 2...M \tag{1}$$

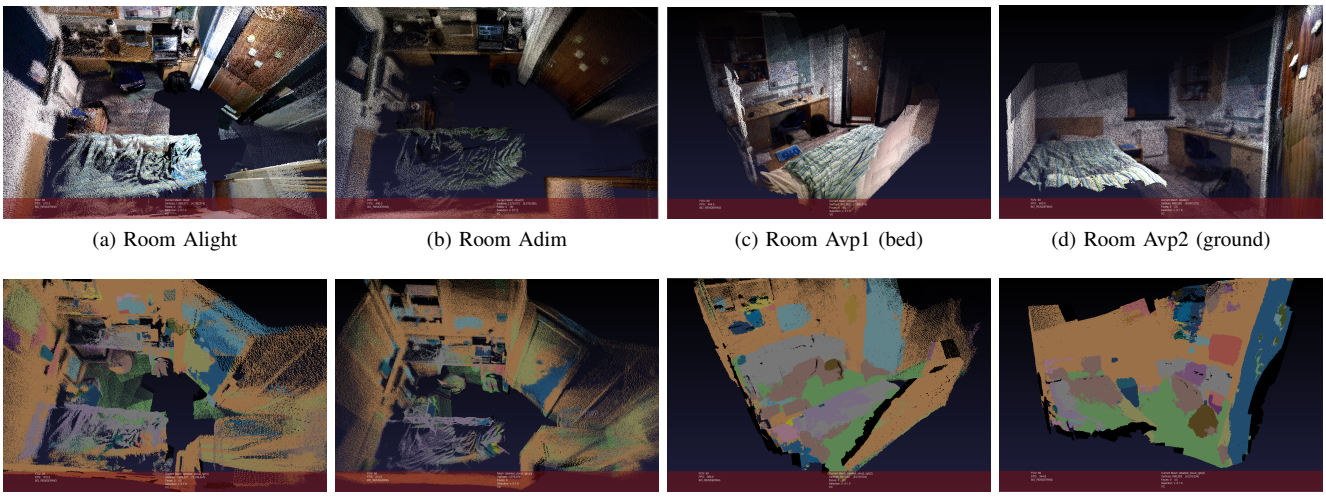
$$z_k^{SWP} = \{z_i^{SWP}\}, i = 1, 2...M \tag{2}$$

where z_i^{SW} and z_i^{SWP} are the tf-idf-based similarity measures (plus a cosine similarity measure) between the current place k and each of the rest of places $i = 1, 2, ...M$ computed using the BoSW and BoSWP descriptors respectively. Each tf-idf score is calculated as shown in equation 3, which is a modified tf-idf measure from the work of Angeli et al.[8]:

$$tf - idf = \frac{n_{wi}}{n_i} log \frac{N}{n_w} \tag{3}$$

In our modified tf-idf score, n_{wi} is the number of occurrences of word w (i.e., number of points with semantic label w) at place i; n_i is the total number of words at place i; n_w is the total number of occurrences of word w of all places i = 1...M; and N is the total number of words. It can be seen from equation 3 that tf-idf score is the product of two frequency terms, namely, the frequency of a word in a place by the inverse frequency of a word of all places. As claimed by Angeli[8], this results in giving increased emphasis to words seen frequently in a small number of places, and penalizing common words that are seen across many places. Once BoSW and BoSWP are weighted by tf-idf scores, similarity measures between different places are computed by taking cosine similarity of the associated tf-idf-weighted scores for different places, BoSW and BoSWP respectively. This leads to the solution sets as shown in equation 1 and 2.

Taking cues from Kejriwal et al.[2], we computed individual likelihood for BoSW and BoSWP descriptors respectively



(a) Room Alight (b) Room Adim (c) Room Avp1 (bed) (d) Room Avp2 (ground)

(e) Room Alight (f) Room Adim (g) Room Avp1 (bed) (h) Room Avp2 (ground)

Fig. 4: Sample pointclouds from the database used to test the algorithm. The top row displays the registered pointclouds that the SLAM algorithm (RTAB-Map) outputs after processing the RGB-D image sequences. The bottom row displays the resulting pointcloud after the segmentation fusion step discussed in Section IV-C.

with equations 4 and 5:

$$L_{SW}(X=Y) = \begin{cases} \frac{z_i^{SW} - \sigma_{SW}}{\mu_{SW}}, & \text{if } z_i^{SW} \geq \sigma_{SW} + \mu_{SW} \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

$$L_{SWP}(X=Y) = \begin{cases} \frac{z_i^{SWP} - \sigma_{SWP}}{\mu_{SWP}}, & \text{if} z_i^{SWP} \geq \sigma_{SWP} + \mu_{SWP} \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where μ_{SW}, μ_{SWP}, σ_{SW}, σ_{SWP} are means and standard deviation of similarity measures for BoSW and BoSWP respectively, and i corresponds to all different places other than the one being compared. The observation likelihood for places having low similarity with the current place are multiplied by 1. This means that the posterior probability would remain the same in the Bayesian update. The joint likelihood of BoSW and BoSWP for each place = i is calculated as equation 6. Every place has a joint likelihood with any of the other places, where the place with the highest joint likelihood corresponds to the best-matched place.

$$L(X=Y) = L_{SW}(X=Y)L_{SWP}(X=Y) \tag{6}$$

Finally, we define a confidence threshold where the nearest neighbor ratio of the highest and second highest joint likelihood should exceed. This is to avoid ambiguity in cases where a place may have multiple matches all with very similar likelihood. In other words, the best match is only considered valid if it is a significantly better match compared to the second best one.

IV. DATASET CREATION

To test the proposed algorithm and more specifically to prove the underlying concept of the properties of DSL, we created a small dataset comprised of single room scans. The different scans have been carried out under different lighting condition and viewpoints changes to test the robustness of the algorithm against these factors. This simple dataset was made to prove the underlying idea that DSL can achieve lighting and viewpoint invariant place description, in contrast to standard dataset testing that aims to benchmark algorithms rather than asserting its properties. Figure 5 shows representative frames of the scans we made using an Asus Xtion sensor to showcase the different scanning conditions and the different appearance of the rooms.

Fig. 5: Dataset overview of the different scans under various lighting changes and viewpoints

A. Scan registration

To register all the RGB-D frames recorded during the scan into a single pointcloud, RTAB-Map (Real-Time Appearance-Based Mapping)[9] is used. RTAB-Map is an RGB-D Graph-Based SLAM algorithm using an incremental appearance-based loop closure detector. The specific loop closure algorithm uses SURF for local feature extraction, and consequently is not robust to large viewpoint or non-affine lighting changes. Consequently, the constraints added to the graph from detected loop closures are used to generate maps. Specifically for each complete scan, RTAB-Map

generates a globally registered RGB pointcloud and a series of camera poses corresponding to the moving trajectory that the camera takes during scanning. Each camera pose has an associated RGB frame (i.e., the RGB image captured by the camera at that camera pose). These three data, namely global pointcloud, camera poses and associated RGB images from RTAB-Map are essential for the rest of the work of our project.

B. Semantic segmentation

SegNet[1] is a Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise segmentation. It learns to predict pixel-wise class labels from supervised learning. It is trained with the dataset of both road scenes and SUN RGB-D. SegNet was primarily motivated by scene understanding applications. Hence, it is designed to be efficient both in terms of memory and computational time during inference. It is also significantly smaller in the number of trainable parameters than other competing architectures and can be trained end-to-end using stochastic gradient descent. Whereas, The performance of SegNet is well correlated with the size of the classes in indoor scenes.

We use a pre-trained SegNet model trained with SUN RGB-D indoor dataset for semantic labeling. SegNet is a general purpose segmentation model which doesn't take into account the depth information. The network takes raw RGB images and densely segment them into 37 indoor scene classes, including wall, floor, ceiling, table, chair, sofa ...etc. In our algorithm we extract segmented results from the second-last layer of SegNet, which outputs the classification scores of all pixels with respect to all semantic classes. This resulting output is of size $480 \times 640 \times 37$, where the dimensions correspond to the image height, width and the number of semantic classes respectively. In order to convert the classification score outputted by SegNet into a probability distribution, for every pixel we algebraically make the smallest score across all 37 classes to 0, adjust the rest accordingly, and then normalize all 37 adjusted scores to obtain the probability distribution.

C. Merging multiple frame segmentations

It has been established that dense semantic information has huge advantages for intelligent robot performance in high level tasks and navigation. Hence, to be able to utilize that information in maps , we need to fuse the segmented results from different 2-D frames given by SegNet into the 3-D global pointcloud. We partially implement fusing DSL based on the Bayesian approach proposed by McCormat et al.[7]. Points in the global pointcloud are initialized with a uniform probability distribution over semantic classes. By using the camera pose from RTAB-Map associated to each RGB frame and the intrinsic matrix from the ASUS Xtion, we project the semantic probability distribution derived from SegNet into the global pointcloud. In details, firstly the frustum culling method is applied for every camera pose in sequence. This would filter out all points in the global pointcloud except for the ones within the camera's field of view at each camera pose (i.e., only points seen by the camera at that camera pose are kept). This is followed by the projection of 3-D points (the points associated with a camera pose) into the pixel space. Once the association between 3-D points and 2-D pixel coordinates are established, the semantic probability distribution at each pixel is back-projected to the associated 3-D point. This back-projected semantic distribution is then multiplied by the distribution a 3-D point currently has (and then normalized), updating the semantic distribution in a Bayesian way. Such Bayesian approach allows updating any 3-D point's semantic distribution from multiple observations. Subsequent observations of the same point will eventually converge and lead to a more confident labeling. Equation 7 below shows the mathematical model of projecting a 3-D coordinate onto its associated 2-D pixel coordinate. Back-projection from 2-D to 3-D is simply the inverse operation of equation 7.

$$^{I}p = {}^{I}P_{C} \cdot {}^{C}K_{S} \cdot {}^{S}A_{W} \cdot {}^{W}p \tag{7}$$

where ${}^{W}p$ is the point's 3-D world coordinate; ${}^{S}A_{W}$ is a transform converting from world to SLAM (RTAB-Map's map reference); ${}^{C}K_{S}$ is the camera extrinsic matrix; ${}^{I}P_{C}$ is the camera intrinsic matrix, and ${}^{I}p$ are the 2-D pixel coordinates associated with ${}^{W}p$.

V. EXPERIMENTAL RESULTS

In this section, we discuss about the experimental results obtained from matching the acquired scans to evaluate the performance of the proposed method.

	Correct	Incorrect	Accuracy
BoSW	6	3	66,67%
BoSWP	2	7	22,22%
BoSW + BoSWP	7	2	77,78%

TABLE I: Results from the matching of the different scans in the dataset considering the likelihood from: only Bag of Semantic Words, BoSWP and the joint likelihood. The two varied viewpoint scans from Room A have been excluded due to the low quality of the segmentation with this scanning procedure.

Table I shows the matching results of the proposed descriptors individually and jointly. The results suggest that indeed dense semantic labels can be successfully used to achieve lighting invariant place recognition. The results with respect to only BoSW are slightly improved by the addition of the word pairs, proving that spatial adjacency is a key aspect to further improve the results. The scans taken to test the viewpoint invariance hypothesis have very different semantic labeling that lead to very poor matching results. The viewpoint invariance hypothesis will have to be tested using a different scanning procedure that offers a more robust segmentation by SegNet or be tested with another semantic segmentation algorithm that doesn't have this problem altogether.

The two misclassified samples, room D and Clight, are mismatched between them due to the BoSW likelihood being

high while the BoSWP likelihood remains low. This suggests that the BoSW likelihood has a big perceptual aliasing problem, since neither of the two rooms look alike in terms of geometry or semantic information. This highlights the need for either, a descriptor less prone to perceptual aliasing or a more restrictive matching procedure that penalizes lower likelihoods. The way that the likelihood is computed currently clips off low likelihood values to one with the aim of not penalizing a high likelihood value from the other. This decision was made since neither of BoSW or BoSWP are highly robust on its own, hence an incorrect likelihood from any of the two shouldn't influence a positive result from the other.

The two only samples matched correctly by BoSWP are the ones corresponding to the Kitchen scans. Kitchen 1 specifically benefited from the addition of BoSWP, being incorrectly matched by BoSW and correctly matched in the joint result.

VI. CONCLUSION AND FUTURE WORK

In this project, indoor place recognition under different lighting and viewpoint conditions is carried out utilizing DSL features. Our proposed algorithm incorporates an available deep learning network for semantic segmentation and a loop-closing-enabled 3-D mapping tool for minimal-drift scans of indoor scenes. Inspired by related literature, our method of fusing DSL information from multiple 2-D views into the global 3-D scene via Bayesian updating framework not only creates indoor maps with semantic representations, but also improves local mapping consistency compared to direct projections. Based on analysis of a place's semantic statistics and spatial adjacency of objects, we proposed modified place descriptors (BoSW and BoSWP). The descriptors were tested on a number of indoor scenes under different lighting conditions and were observed to produce consistent recognition results. On the other hand, scans obtained with drastic viewpoint changes were tested but the proposed descriptors did not correctly identify places. It was observed that the same places scanned from different viewpoints presented inconsistent DSL. This implies that our proposed algorithm is bounded in its performance by the accuracy of the semantic segmentation.

A number of future works are to be carried out to improve place recognition using DSL features. First of all, to further improve mapping consistency, McCormac et al. [7] made use of the geometry of the 3-D map to regularize semantic prediction after fusing DSL into 3-D pointcloud. This post-fusion enhancement was not implemented in our current development and will be carried out. Moreover, in addition to incremental Bayesian update to fuse and improve semantic features consistency, we would like to include a Gaussian weighting scheme based on expected objects dimension to also improve single-frame semantic labeling. Furthermore, the place descriptor based on spatial adjacency histogram (BoSWP) can be extended to a graph-based representation to fully describe spatial relationship of objects in the scene. In such graph-based approach, each node in the graph corresponds to an object with properties (i.e., semantic label and color statistics), and the edge linking two nodes describes spatial relationship in terms of distance and orientation between any two objects. A more suitable matching scheme dedicated to this graph matching will also be used to measure similarity score. We also plan to conduct comparisons with state-of-the-art algorithms by using standard datasets that include normal, varied lighting and viewpoint conditions to benchmark the improved algorithm using semantic information for place recognition.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[2] N. Kejriwal, S. Kumar, and T. Shibata, "High performance loop closure detection using bag of word pairs," *Robotics and Autonomous Systems*, vol. 77, no. Supplement C, pp. 55 – 65, 2016.

[3] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[4] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "Fab-map+ ratslam: Appearance-based slam for multiple times of day," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 3507–3512, IEEE, 2010.

[5] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," *CoRR*, vol. abs/1504.05241, 2015.

[6] S. Cascianelli, G. Costante, E. Bellocchio, P. Valigi, M. L. Fravolini, and T. A. Ciarfuglia, "Robust visual semi-semantic loop closure detection by a covisibility graph and cnn features," *Robotics and Autonomous Systems*, vol. 92, no. Supplement C, pp. 53 – 65, 2017.

[7] J. McCormac, A. Handa, A. J. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," *CoRR*, vol. abs/1609.05130, 2016.

[8] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words.," *IEEE Trans. Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.

[9] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 734–745, 2013.