RGB-D SLAM
Using Attention Guided Frame Association
Ali Caglayan1 *, Nevrez Imamoglu1 *, Oguzhan Guclu2 *, Ali Osman Serhatoglu3 *, Weimin Wang4, Ahmet Burak Can3 and Ryosuke Nakamura1
1National Institute of Advanced Industrial Science and Technology, Tokyo JAPAN
Email: {ali.caglayan, nevrez.imamoglu, r.nakamura}@aist.go.jp
2Sahibinden, Istanbul, Turkey
Email: guclu.oguzhan@outlook.com
3Hacettepe University, Ankara TURKEY
Email: {ao.serhatoglu, abc}@cs.hacettepe.edu.tr
4Dalian University of Technology, Dalian, China
Email: wangweimin@dlut.edu.cn

Abstract—Deep learning models as an emerging topic have shown great progress in various fields. Especially, visualization tools such as class activation mapping methods provided visual explanation on the reasoning of convolutional neural networks (CNNs). By using the gradients of the network layers, it is possible to demonstrate where the networks pay attention during a specific image recognition task. Moreover, these gradients can be integrated with CNN features for localizing more generalized task dependent attentive (salient) objects in scenes. Despite this progress, there is not much explicit usage of this gradient (network attention) information to integrate with CNN representations for object semantics. This can be very useful for visual tasks such as simultaneous localization and mapping (SLAM) where CNN representations of spatially attentive object locations may lead to improved performance. Therefore, in this work, we propose the use of task specific network attention for RGB-D indoor SLAM. To do so, we integrate layer-wise object attention information (layer gradients) with CNN layer representations to improve frame association performance in a state-of-the-art RGB-D indoor SLAM method. Experiments show promising initial results with improved performance.

I. INTRODUCTION

Advances in deep learning have provided great results in many computer vision tasks such as image/scene recognition [1, 2], object recognition/detection [3, 4], image-to-image generation (e.g. semantic segmentation [5], super resolution images [6], style transfer[7], etc.). Deep learning approaches have also become one of the core components of machine vision applications in autonomous systems such as simultaneous localization and mapping [8], which is a very crucial part of autonomous robots or self-driving cars. As stated in [9], however, there is still need for deep learning based SLAM research, especially for advanced tasks such as with geometric task or frame association. Therefore, in [9], CNN representations (features) of images extracted from a pre-trained network are used to handle loop closure detection in a SLAM framework, which yields improved performance compared with the state-of-the-art models on TUM RGB-D benchmark [10].

On the other hand, methods such as class activation mapping (CAM) [11] or gradient based CAM (Grad-CAM) [12] have enabled researchers to visualize networks attention area that can explain which visual area of a scene is more effective to the decision given by the trained network for a given input image. It would be expected that the higher-level network semantic features in the associated area would be more important to the network task. This also helps to make models for weakly-supervised detection/segmentation of objects [13] without the requirement of large-scale training data with pixel-wise labels of region of interests.

In addition, in [14], CAM [11] visualization approach is directly integrated as a sub-network (i.e. attention branch network), which aims to learn network attention map (or CAM) and modulate CNN features to improve recognition task of the network as the classification branch. This requires additional trainable parameters for the attention branch to the network model. In contrary, gradient based approaches such as Grad-CAM [12] can also obtain network attention maps without the need of additional parameters. For example, inspired by [12] and [13], in [15], authors enhance class-specific weakly supervised object detection to extend a more generalized task dependent (transfer-learning) salient object detection model. This is achieved by integrating gradients with CNN features for localizing attentive regions for all possible distinct objects, which can be beneficial to many visual tasks. For example, this gradient based object attention model in [15] is used for obtaining more user centric 360° video streaming for VR systems in [16]. Though, deep features are only used to obtain saliency (attention) maps that are used for bit-rate allocation of the tiles on 360° frames defining their importance.

Despite these progresses, there is not much explicit usage of these gradient (network attention) information (except visualization) to integrate with CNN representations for object semantics in complex vision models such as SLAM. In fact, without the need of any extra fine-tuning or training, transfer-learning based gradients (cues of network attention for distinctive objects) may guide to suppress background
In this work, we propose and investigate the use of task specific network attention to improve RGB-D indoor mapping (see Fig. 1). To do so, we integrate CNN representations of a semantic deep layer with gradients or layer-wise object attention maps obtained by a pre-trained network of the ImageNet [1] as in [15]. Thus, we aim at obtaining representations focused on objects’ attention in the network with suppressed background. Then, these guided features are utilized to improve loop closure detection performance of the RGB-D SLAM method in [9]. Moreover, in this work, attention mechanism is investigated for only frame associations for a better loop closure detection using color images; however, it is possible to apply the similar approach in motion estimation for a more efficient keyframe/keypoint selection. Experiments show promising initial results with improved mapping performance for color based image association by comparing high-level encoded features.

II. PROPOSED METHOD

The SLAM module employed in [9] is a graph-based system which adopts a feature based approach for odometry estimation and a deep feature indexing mechanism for loop closure detection. A pose graph is built by inserting nodes for each incoming frame and establishing odometry and loop closing edges between related nodes. The environment map is constructed by optimizing the graph and then projecting point clouds at each pose into a coordinate frame.

For the camera movements, the system extracts keypoints on the corresponding color frames, locates them in 3D by utilizing related depth frames, and computes the transformation through applying RANSAC on the keypoint matches. This process is carried out between consecutive frames to estimate the odometry. For loop closure detection, a deep feature mechanism based on a task specific network attention model (see Section II-A) is employed. The difference of this work compared with [9] is in the deep feature extraction module shown in Fig. 1. This work does not utilize CNN features from VGG network directly. Instead, deep features extracted from layers are modulated by the gradients, representing the object attention information, to suppress the background information.
Then these attention guided features are passed through RNNs for encoded compact representations. In this manner, deep features are extracted from keyframes and indexed in a priority search k-means tree [17] through being handled as a point in high dimensional space. The search tree is built incrementally during the navigation by inserting deep features of every determined keyframe. For each incoming frame, the system performs a loop closure search in the tree according to the deep feature similarity. A group of candidates are obtained firstly. Then, an adaptive thresholding approach is applied to eliminate outliers. The motion estimation process (same as in the odometry step) is applied between each of the final candidates and the incoming frame, and loop closures are determined according to the transformation quality.

The loop closure search process is vital for the map accuracy since an incorrect loop closure detection might cause a graph optimization failure and thus a wrongly built map. Simple yet a beneficial approach for improving scene representation based on gradients for salient objects (i.e. network attention) can boost performance (e.g. up to 10 to 20 cm in some samples of TUM RGB-D benchmark dataset [10]).

A. Deep Features for Frame Association

1) Object Features Guided CNN Features: Deep feature extraction module in Figure 7 provides the attention guided compact representation to be used in the loop closure detection process. Proposed attention guided representation module takes advantage of the task specific salient object detection model using forward and backward features for the ImageNet pre-trained VGG network as proposed in [15]. In this work, we focus only on the deep representations at CNN Block 5 (see Figure 7), which have better semantic representation compared with the other layers [3], [18].

Unlike setting initial gradients for a specific class to 1 and others to 0 such as Grad-CAM [12] or distinct class saliency [13], this approach [15] takes advantage of the actual object class prediction scores obtained from the softmax output of VGG network. The class prediction scores are used as initial gradients to start back-propagation for computing object saliency values (gradients) for all possible attentive objects at a desired level of the network \( L_l \) independent of the object class. For the predicted class scores, gradients of a selected layer \( l \) can be formulated as in Eq. 1:

\[
G_l = \frac{\partial S}{\partial L_l}
\]  

where \( G_l \) is the derivative of object scores \( S \) with respect to the layer at \( L_l \) feature activation [13]. During back-propagation in the backward process, we use partially guided back-propagation between separated blocks at each max-pooling layer for efficiency. This is done by setting all negative gradients only at these transitions between blocks [15]. This is different from the guided back-propagation used in [13] that sets all negative gradients at any layer to 0. After obtaining the gradient at a selected layer, extraction of attention guided feature \( F_l \) is given as below:

\[
F_l = \delta(L_l, G_l)
\]  

where \( \delta \) is the fusion function for the feed-forward CNN layer features \( L_l \) with the gradient features \( G_l \) representing the attentive object regions. For a given layer \( l \), we explore the different ways to integrate object attentive features \( G_l \) to modulate forward features \( L_l \) to suppress background scene. So we define different fusion strategies for using \( G_l \) such as directly using normalized gradient tensor as in Eq. 3 and 4 or creating single normalized object saliency map by channel-wise summation of gradient tensor as in Eq. 5 and 6. We term these attention functions in Eq. 3, 4, 5, and 6 as mult, exp, sumdim, and exp_sumdim, respectively.

\[
\delta(L_l, G_l) = L_l \odot N(G_l)
\]  

(3)

\[
\delta(L_l, G_l) = L_l \odot e^{N(G_l)}
\]  

(4)

\[
\delta(L_l, G_l) = L_l \odot N\left( \sum_i N(G_{ij}) \right)
\]  

(5)

\[
\delta(L_l, G_l) = L_l \odot e^{N(\sum_i N(G_{ij}))}
\]  

(6)

Here, \( \odot \) represents the Hadamard product and \( N(\cdot) \) indicates the normalization operation that scales \( G \) to the gradient range of [0-1] as the attention guide for \( L \). Unlike [15], we normalize the gradients to suppress only the activations related to background clutter. In this way, we extract attention guided features, where activation values related to the salient objects are more dominant with respect to the background of the scene.

2) Random Recursive Neural Network for Feature Encoding: After obtaining object attention guided CNN features from the optimum CNN block 5 of the VGG network (we term these as L5 as in [18]), next step is to encode these representations into a more compact space. For frame to frame comparison of scene content, direct usage of these representations may hinder the performance of SLAM due to large feature space (i.e. curse of dimensionality). Therefore, we employ random recursive neural networks (RNNs) [19] to pool CNN features in a lower dimensional space while providing a compact and separable feature space as in [9]. However, unlike [9], we apply the reshaping process after passing the CNN activation maps through an average pooling process. We apply average pooling over the high number of activation maps that extracted from the deep VGG level of L5 in order to fit the previous more straightforward AlexNet model in [9]. Therefore, we first pool every two maps of the extracted activations into one map by a averaging pixel values simply. This results in halving the number of maps (i.e. 7 x 7 x 512 to 7 x 7 x 256). Then, we reshape these activations into the dimension of 14 x 14 x 64 to feed the RNNs. An RNN encodes CNN features into a lower dimensional space on a graph structure by applying the same merging adjacent vectors into a parent vector with tied weights recursively [19]. Then, the parent vector is passed through a nonlinear activation function (i.e. tanh). We use the implementation of [18], which is a one-level structured...
Figure 2. Attention guided model accuracy performances over the baseline [9] in terms of RMS-ATE (root mean square of absolute trajectory error in meters) on the fr1 sequences.

RNN with a single parent vector. Each RNN produces $K$-dimensional feature vector (i.e. $K = 64$). As in [9], we apply 16 RNNs and this results in a total of $64 \times 16 = 1024$ dimensional feature vector.

III. EXPERIMENTS

We evaluated the performance of the proposed model with the popular TUM RGB-D dataset [10]. In order to assess both medium and large scale mapping performance, the fr1 and fr2 sequences of the dataset were employed. The fr2 sequences contain scenes of a larger environment and have much more difficult environmental conditions than fr1.

Figure 2 shows accuracy performance of the attentive system using the defined different attention functions together with the baseline results on the fr1 sequences. The average results on the fr1 sequences are 0.0364, 0.0369, 0.0367, 0.0363, and 0.0366 for the baseline, sumdim, exp, exp_sumdim, and mult, respectively. As these results and the graph indicate, the success of the object attentive features on the baseline is not clear and they show very similar performance. This is due to the fact that the margin of error is quite low in the smaller scaled fr1 sequences, the object attentive based contribution is somewhat vague on these data. For example, if the sequence of sample data is around one specific object, it is neither easy nor feasible for the network to distinguish foreground object and background regions using the proposed object attentive gradients. Because, if the focus of camera/image frame is too close to a single object, low-level features might have better reliability than the more semantic object attentive features. It can be seen that attention guide in fr1 sequences does not improve the baseline. However, it should be noted that the results of fr1 sequences are very close to the ground-truth results in both baseline method and the proposed attention guide integration models.

On the other hand, for the larger fr2 sequences, we can see the contribution of the object attentive features more precisely. This also indicates that an attention-based SLAM approach might be important in large-scale outdoor mapping methods by focusing on different object properties in an environment.

Table I shows the obtained object attentive results with the baseline method that on TUM fr2 sequences. As can be seen from the average results, all of our attentive objects guided model approaches have significantly improved the results by decreasing the RMS-ATE compared to the baseline that utilizes only convolution maps based features. The observed drift errors are close to 10 cm - 35 cm interval, which are acceptable for these highly challenging sequences recorded in a large indoor environment. Meanwhile, our ablative studies on different strategies on the fusion of forward (feed-forward process outputs at layer 5) and backward (object attentive gradients) features show that the mult attention fusion function brings consistent accuracy improvements over all the other approaches. Figure 3 shows some sample trajectories using our proposed mult object attentive model on the fr1 plant, fr2 large_no_loop, fr2 pioneer_slam, and fr2 pioneer_slam3, respectively. The proposed model minimizes the RMS-ATE errors and obtains quite consistent trajectory maps comparing to the ground-truth results.

IV. CONCLUSION

In this paper, we propose and investigate gradient based object attentive loop closure detection in an indoor RGB-D
mapping system. To the best of our knowledge, this paper is the first attempt to attention guided features in a SLAM system by modulating forward representation with object attentive gradients. The experimental results are very promising for further improvements such as using an eye-fixation trained network instead of a classifier network. Especially the impressive results obtained on the large-scale fr2 sequences indicate that our approach might be prominent in outdoor mapping as well. Moreover, attention networks feature can be fully investigated not only for frame association but also for key-point detection, key-frame selection, descriptor, and etc. Finally, the use of only RGB frames has been investigated in this paper. Using RGB and depth data modalities together in a multi-modal approach might further improve the results.

REFERENCES

[1] O. Russakovsky*, J. Deng*, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei*, “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, pp. 211–252, 2015.

[2] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40.6, pp. 1452–1464, 2017.

[3] A. Caglayan and A. B. Can, “Exploiting multi-layer features using a cnn-rnn approach for rgb-d object recognition,” in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, September 2018.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, June 2010.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[6] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[7] F. Luan, S. Paris, E. Shechtman, and K. Bala, “Deep photo style transfer,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[8] R. Mur-Artal and J. D. Tardos, “Orb-slam2: an open-source slam system for monocular, stereo and rgb-d cameras,” IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255–1262, 2017.

[9] O. Guclu, A. Caglayan, and A. B. Can, “Rgb-d indoor mapping using deep features,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

[10] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in Proc. of the International Conference on Intelligent Robot Systems (IROS), Oct. 2012.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[13] W. Shimoda and K. Yama, “Distinct class-specific saliency maps for weakly supervised semantic segmentation,” in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, October 2016.

[14] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, “Attention branch network: Learning of attention mechanism for visual explanation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[15] N. Imamoglu, C. Zhang, W. Shmo, Y. Fang, and B. Shi, “Salience detection by forward and backward cues in deep cnn,” in 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 430–434.

[16] C. Ozcinar, N. Imamoglu, W. Wang, and A. Smolic, “Delivery of omnidirectional video using saliency prediction and optimal bitrate allocation,” Signal, Image and Video Processing, pp. 1863–1711, September 2020.

[17] M. Muja and D. G. Lowe, “Scalable nearest neighbor algorithms for high dimensional data,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 11, pp. 2227–2240, 2014.

[18] A. Caglayan, N. Imamoglu, A. B. Can, and R. Nakamura, “When cnns meet random rns: Towards multi-level analysis for rgb-d object and scene recognition,” Computer Vision and Image Understanding, p. 103373, 2022.

[19] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, “Convolutional-recurrent deep learning for 3d object classification,” in Advances in Neural Information Processing Systems, 2012, pp. 656–664.

Figure 3. Comparison of estimated trajectories based on our multi attentive object model with the ground-truths on the fr1_plant, fr2_large_no_loop, fr2_pioneer_slam, fr2_pioneer_slam3 sequences respectively.