IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 6, 2019, accepted February 10, 2019, date of publication February 13, 2019, date of current version March 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899049

A Coarse to Fine Indoor Visual Localization Method Using Environmental Semantic Information

WEI ZHANG, GUOLIANG LIU, (Member, IEEE), AND GUOHUI TIAN, (Member, IEEE)
School of Control Science and Engineering, Shandong University, Jinan 250061, China

Corresponding author: Guoliang Liu (liuguoliang@sdu.edu.cn)

This work was supported in part by the the National Key R&D Program of China under Grant 2018YFB1306504, in part by the National Natural Science Foundation of China under Grant 61603213 and Grant 91748115, in part by the China Postdoctoral Science Foundation under Grant 2016M590635, in part by the Young Scholars Program of Shandong University under Grant 2018WLJH71, and in part by the Taishan Scholars Program of Shandong Province.

ABSTRACT In this paper, we focus on the camera localization problem using visual semantic information. In contrast to the state of the artworks which often use visual features to do localization, we here propose a coarse to a fine mechanism to localize the camera position. First, a semantic database including object information around the target environment is constructed using a deep learning method. Second, for the coarse step of the visual localization, we match class attributes of objects in the current frame to the object database and find candidate frames that have similar objects. Third, the most similar candidate frame to the current frame is selected by CNN features. For the fine step of localization, the final pose of the camera can be estimated using feature matching with semantic information. Compared to the state of the art visual localization methods, the proposed localization method based on semantic information has higher localization accuracy. Furthermore, the proposed framework is not only useful for visual localization, but also useful for other advanced tasks of robot, e.g., loop closing detection, object searching, and task reasoning.

INDEX TERMS Visual localization, semantic mapping, CNN features, SLAM.

I. INTRODUCTION

Simultaneous localization and mapping (SLAM) is a process for robots to locate itself and build a map in an unknown environment with laser, camera or other sensors. In recent years, with the rapid development of computer vision methods and the improvement of visual sensors performance, visual SLAM has made a significant progress. Traditional visual SLAM algorithms mainly focus on the precise reconstruction of the environment geometry, which can be used for localization. However, the robot requires more semantic information of the working environment to finish advanced tasks, e.g., find a cup in the kitchen and pouring water into it. The geometry is important, but the semantic is the future, which is a basic and necessary capability for smart service robot. Semantic SLAM refers that the SLAM system not only includes traditional geometric structure information, but also the semantic information such as objects attribution [1], [2]. The advantages of semantic SLAM over traditional SLAM are following [3]: firstly, traditional SLAM algorithms usually assume that the environment is static, but the semantic SLAM can predict the movable property of objects. Secondly, semantic SLAM can achieve intelligent path planning, e.g., the robot can move the objects away in the environment to get a better path. Thirdly, semantic SLAM can improve the storage efficiency and scalability of SLAM system using the knowledge sharing and management of similar objects. The core of semantic SLAM is accurate object recognition in the environment.

Deep learning technology has achieved rapid development in the field of object recognition, which can be used for semantic mapping. Ren *et al.* [4] propose a multi-object detection method based on the Faster R-CNN network. It utilizes a Region Proposal Network (RPN) network for area selection, and combines with the Fast R-CNN network to detect multiple objects. Liu *et al.* [5] introduce a direct target object detection algorithm based on SSD network, which can save the detection time compared to the Faster R-CNN network by removing the RPN for area selection. However, these methods can only obtain a rectangular area which includes the target object and some background information. To estimate a better boundary of the target object,

The associate editor coordinating the review of this manuscript and approving it for publication was Huazhu Fu.

He *et al.* [6] propose an instance segmentation framework based on Mask R-CNN network, which adds a branch layer for predicting the target mask on the basis of Faster R-CNN network, and achieves pixel-level object segmentation.

The improved object recognition and segmentation accuracy using deep learning have attracted many researchers to explore new methods for semantic SLAM [7]. McCormac *et al.* [8] employ a convolutional neural network (CNN) for dense 3D semantic map construction, which uses the Elastic Fusion SLAM [9] algorithm to estimate the egomotion, and predicts the pixel-level object class labels in 2D key image frames based on CNN. These labels can be upgraded using conditional random fields and Bayesian theory according to different views. Li and Belaroussi [10] propose to use the CNN for semi-dense 3D semantic map construction, which utilizes the LSD-SLAM [11] algorithm to estimate the camera pose, and uses the conditional random fields for noise smoothing. These semantic SLAM algorithms consider the deep learning as the segmentation tool to classify the 2D pixels and 3D point clouds in the map, but how to manage these objects and how to use the semantic map is still a undergoing question.

Visual localization as an indispensable part in visual SLAM system has made great progress in recent years. Especially for the rapid development of Deep Learning technology, a number of alternative technologies have been proposed for visual localization. Chen *et al.* [12] use the CNN features to retrieve the most similar image, and then estimate the camera pose with the corresponding ORB feature points. Kendall *et al.* [13] construct a CNN model (PoseNet) to directly regress the camera pose with RGB input. Guo *et al.* [14] use the pose predicted by PoseNet as the initial result, and use LSTM as a temporal filter to refine the pose. Taira *et al.* [15] retrieval the candidate images with NetVLAD network, and estimate the camera pose with densely matching. However, The methods mentioned above have not considered the semantic information of images to improve the localization accuracy and efficiency.

In this paper, we propose a deep learning based visual semantic database construction method, and introduce a coarse to fine visual localization method using semantic information. The main contributions of this paper can be concluded as follows:

(1) We construct a semantic database based on objects information, which is helpful for localization.

(2) Our method is training free for different scenarios, because we don't need to train a unique model for specific scene.

(3) The proposed coarse to fine mechanism can increase the localization accuracy and efficiency with the objects information.

The rest of this paper is organized as follows: the proposed ideas are introduced in Section II, and the experimental results are shown in Section III. Finally, we conclude the paper and discuss future works in Section IV.

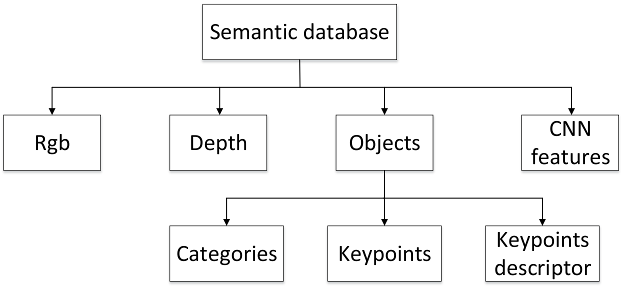


FIGURE 1. The structure of visual semantic database.

II. THE PROPOSED METHOD

This paper constructs the visual semantic database based on the Mask R-CNN network, and uses MySQL to manage the semantic database. The structure of visual semantic database in this paper is shown in Fig. 1. The semantic database includes keyframes with corresponding RGB image, depth image, CNN features, and recognized object information with corresponding keypoints, descriptors and categories.

The algorithm of this paper includes two stages: the construction of visual semantic database and the global visual localization. At the phase of the first stage, we use the Mask R-CNN network to extract the category and position information of objects in the key frames. The keypoints and descriptors corresponding to the recognized objects are extracted using the SURF algorithm. Besides, we also extract the CNN features of the rgb image for scene classification. The CNN features and the objects information are then stored in the visual semantic database. The second stage is the global visual localization, whose flow chart can be seen in Fig. 2. Given an image from a camera, the localization algorithm obtains the CNN image features, detects and recognizes the objects with the Mask R-CNN network, and extracts the SURF features for each object. We then conduct a two-layer filtering mechanism to find out candidate key frames in the semantic database: the object attributes are first used to select the candidate key frames, and the CNN features are then used to select the most similar key frame. Finally, Bundle Adjustment(BA) can be used for estimating the transform matrix between the query image and candidate frames in the semantic database. One of the advantages of our algorithm is that many irrelevant frames in the database can be excluded with the proposed two-layer mechanism, which can improve the matching efficiency.

A. OBJECT DETECTION BASED ON MASK R-CNN NETWORK

The deep learning has showed impressive performance for object detection and segmentation. The Faster R-CNN network proposed by Ren *et al.* [4] and the SSD network proposed by Liu *et al.* [5] can find the objects using a rectangle box, which also includes background information. To achieve better boundaries of objects, we use the Mask R-CNN proposed by He *et al.* [6], which can be seen in Fig. 3.

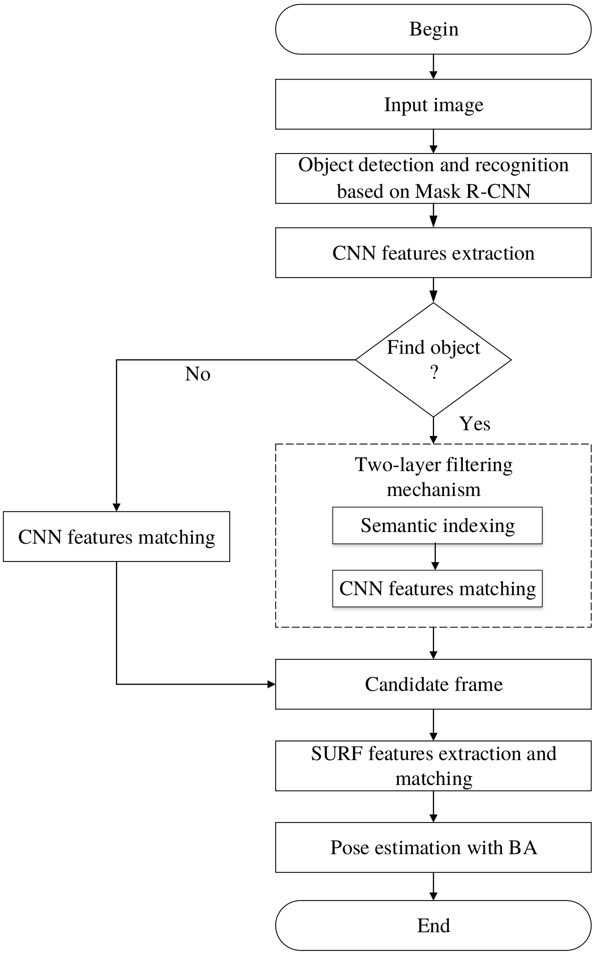


FIGURE 2. **The flow chart of global visual localization based on semantic database.**

The Mask R-CNN network basically has following modules: the Backbone network are used to extract the high-level features of the input image, the candidate regions of objects are acquired using a RPN network, the RoiAlign layer normalizes the features of candidate regions to fixed size for the subsequent classification recognition. There are three types loss: the classification loss L_{cls} to identify the target category, the detection loss L_{box} to regress the square box for the target region, and the segmentation loss L_{mask} to identify the target at the pixel-level. The final loss function is

$$L = L_{cls} + L_{box} + L_{mask}. \tag{1}$$

By minimizing the loss function using the backpropagation algorithm, the Mask R-CNN network can obtain the target category, the square box and the mask simultaneously.

B. DEEP FEATURES EXTRACTION BY CNN

The pose estimation accuracy mainly depends on the similarity of two images, which can be seen as the image retrieval problem. Recent years, deep learning has showed its advantages for solving image retrieval problem [16]. The CNN features in the middle layers exhibit robustness against appearance changes, whereas the features from the top layers are more robust with respect to viewpoint changes. Here we adopt the pretrained resnet50 [17] CNN models on Places dataset, which has shown great performance in place recognition [18]. To handle the viewpoint changes, we use the output of avgpool layer (top layer) in resnet50 as the image features for matching, which has 2048 dimensions, as shown in Fig. 4.

To match two images, the similarity score $score_i$ is calculated using the cosine distance between the L2-normed image features of the queue image ($vector_q$) and the retrieved image ($vector_i$), which is shown as:

$$vector_i = \left\| vector_i^1, vector_i^2, \cdots, vector_i^{2048} \right\|_2, \tag{2}$$

$$vector_q = \left\| vector_q^1, vector_q^2, \cdots, vector_q^{2048} \right\|_2, \tag{3}$$

$$score_i = vector_i * vector_q^T. \tag{4}$$

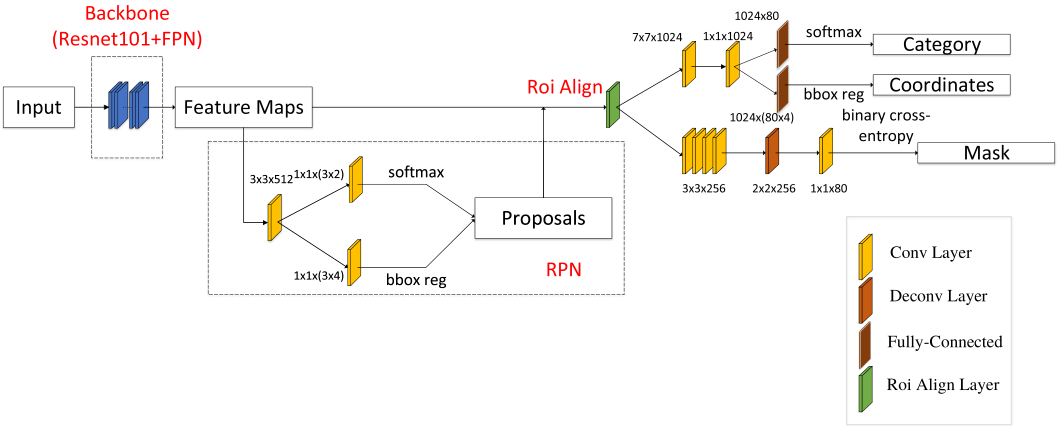
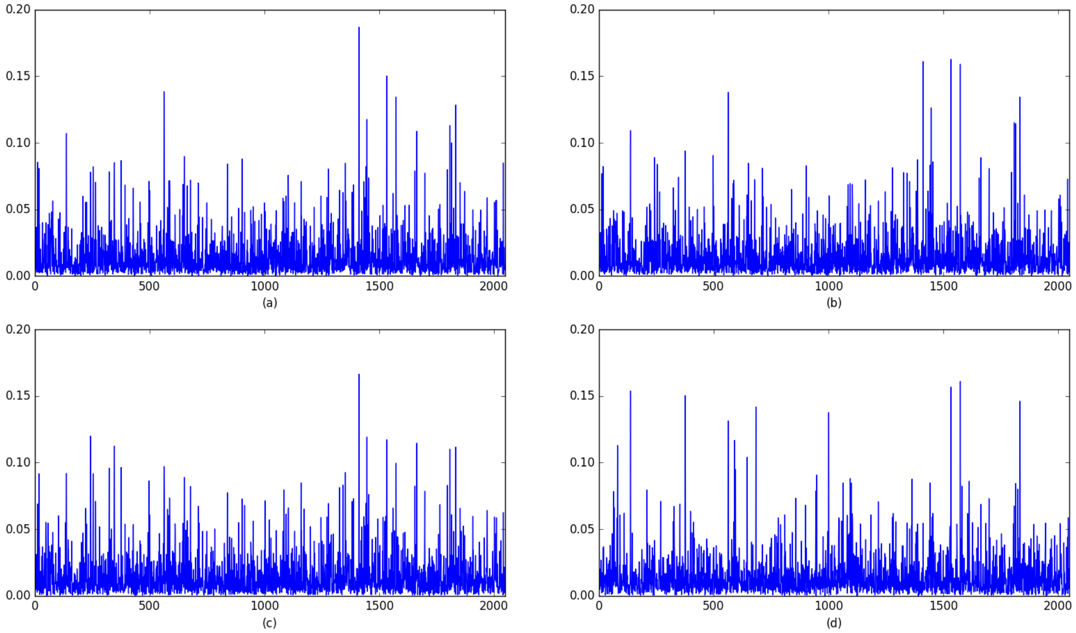


FIGURE 3. **The architecture of Mask R-CNN network.**



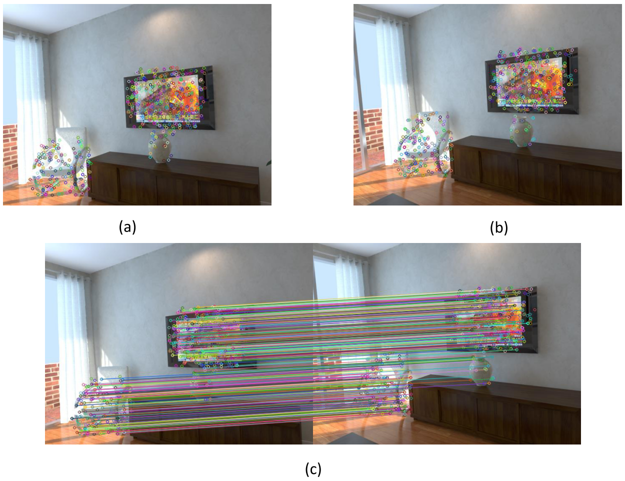
(a) (b) (c) (d)

FIGURE 4. Visualization of CNN features. (a) the features of the queue image; (b) the features of the retrieved image with a top score; (c) the features of retrieved image with a second top score; (d) the features of an unrelated image.

C. OBJECTS FEATURE EXTRACTION AND MATCHING BASED ON SURF

To estimate relative pose between image frames, a feature list corresponding to the object is required. In this work, we use the speed up robust feature (SURF) to detect the feature corners of objects, which is not only faster in speed, but also robust under complex conditions such as scale change and lighting [19]–[21]. The SURF algorithm firstly determines the pre-selected points using the Hessian matrix, which has faster extraction speed than SIFT algorithm. In order to solve the problem of inconsistency in size and resolution of the same object due to distance, the SURF algorithm constructs scale space using different size of Gaussian template. The hessian matrix is used on each scale to obtain candidate extreme points. Then each extreme point is compared with 8 adjacent points on the same scale, 9 points on the upper scale and 9 points on the lower scale. When the extreme point value is the largest or smallest of the 26 points, this extreme point is considered as a true feature point. For the feature descriptor, the SURF algorithm calculates the sum of Haar wavelet responses both in the X direction and Y direction with a radius of $6s$, where s is the scale of the feature. Then the main direction of the feature point is chosen as the longest direction by traversing the whole circular region around the feature point, and the feature descriptor is obtained according to the Haar wavelet response of the main direction.

To accelerate the detection and matching speed, here we only extract the keypoints of objects detected by Mask R-CNN network, and then match them based on the keypoint descriptors. If the Euclidean distance ratio between the nearest neighbor feature point and the second nearest neighbor feature point is smaller than a certain threshold, the pair of the current feature point and the nearest neighbor feature point is added to the feature matching list. The results of feature detection and matching from two images are shown in Fig. 5. Compared to the keypoints extraction and matching from the whole image, our method is more efficient.



(a) (b) (c)

FIGURE 5. Detection and matching of SURF features: (a) and (b) show the detected keypoints in a pair of images; (c) shows the matching result.

D. CAMERA POSE ESTIMATION

We employ the Bundle Adjustment (BA) to estimate the relative pose between the input image and the candidate image. The BA algorithm is also known as the beam adjustment algorithm [22]. The algorithm simultaneously optimizes the

TABLE 1. The experiment datasets from different scenarios of ICL-NUIM and TUM RGB-D.

Scenario	Raw Images	Key Frames	Test Images
living room ’of kt0’	1510	70	1000
living room ’of kt1’	967	36	600
living room ’of kt2’	882	54	500
living room ’of kt3’	1242	101	800
office room ’of kt0’	1510	77	1000
office room ’of kt1’	967	47	600
office room ’of kt2’	882	54	500
office room ’of kt3’	1242	45	800
freiburg1_ plant	1120	171	800
freiburg1_ room	1352	243	1000
freiburg2_ 360_hemisphere	2647	800	2000
freiburg2_ flowerbouquet	2859	323	2000
freiburg2_ pioneer_slam3	2224	253	1800
freiburg3_long_ office_household	2486	246	2000

camera pose and spatial position of feature point with the principle that rays reflected from each feature point can converge to the camera optical center after optimization. The loss function is:

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|z_{ij}-h\left(\xi_i,p_j\right)\right\|^2. \qquad (5)$$

where z_{ij} represents the pixel coordinate of feature point p_j with the camera pose ξ_i, and h represents the observation equation of feature point p_j with the camera pose ξ_i. By minimizing the cost function using nonlinear optimization, BA can estimate accurate camera pose and spatial coordinate of feature points.

III. EXPERIMENT AND ANALYSIS

A. DATA ACQUISITION

Our visual localization method requires the rgb-d images and their associated poses, so we use the ICL-NUIM dataset [23] and the TUM RGB-D dataset [24], [25]. The ICL-NUIM dataset consists of rgb-d images from two indoor scenes, living room and office room. The image size is 640 × 480 and the ground-truth trajectories are offered. The TUM RGB-D dataset includes 89 rgb-d sequences with different camera motions. The images have 640 × 480 resolution and the ground-truth trajectories are captured by a high accuracy motion-caption system. These datasets are employed to verify to performance of our localization method. As illustrated in Table 1, we only use these images that do not have common views as the key frames in our semantic database, whereas other images can be used as test images for localization.

TABLE 2. The detected results on COCO minival dataset.

	Type	IoU	Accuracy
Bounding Box	Average Precision	0.50:0.95	0.354
	Average Precision	0.50	0.546
	Average Precision	0.75	0.385
	Average Recall	0.50:0.95	0.437
Instance Segmentation	Average Precision	0.50:0.95	0.305
	Average Precision	0.50	0.513
	Average Precision	0.75	0.318
	Average Recall	0.50:0.95	0.384

B. NETWORK TRAINING

To recognize the objects in datasets, we fine-tune the Mask R-CNN network that has trained on public COCO dataset. Considering the indoor environment, we select those objects that appear frequently in indoor environment from COCO dataset [26] as the sub-dataset for training, i.e., bottle, chair, potted plant, screen, notebook, mouse, keyboard, cell phone, book, cup and so on. The architecture of Mask R-CNN can be seen from Fig. 3. The Backbone is made up of resnet101 network [17] with FPN (feature pyramid networks) [27], which can extract high-level features of the input image with different scales. The RPN (Region Proposal Network) produces the candidate regions of objects, and the RoiAlign layer normalizes the features of candidate regions to fixed size for the subsequent classification recognition. The fine-tune mechanisms are as following: we first train the classification portion, then we train the FPN plus classification portion with learning rate of 0.001. Finally, all layers of Mask R-CNN network are fine-tuned with learning rate of 0.0001. The results of bounding box and segmentation on COCO minival dataset are reported in Table 2.

C. LOCALIZATION RESULTS AND ANALYSIS

Fig. 6 and Fig. 7 summarize the performance of our coarse to fine localization method. As shown in Fig. 6, more than 80% of the queue images are localized within 2.5 degrees.

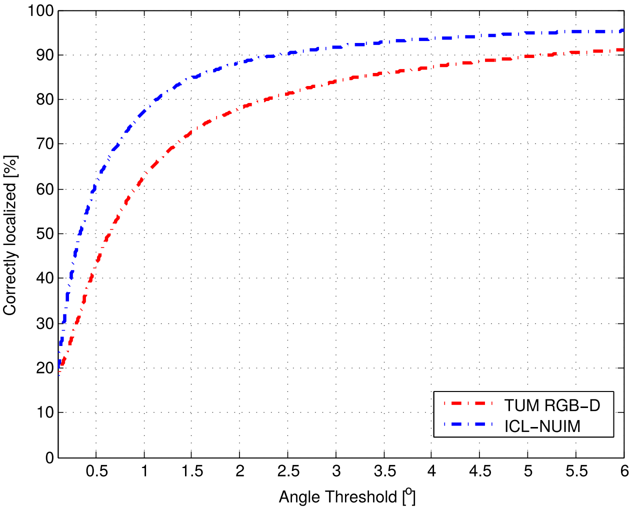


FIGURE 6. Cumulative distribution of angle error.

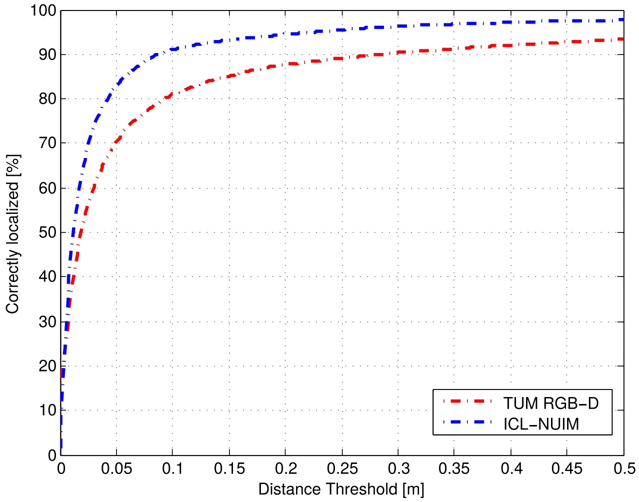


FIGURE 7. Cumulative distribution of location error.

From Fig. 7, we can see that above 90% of the queue images are localized within 0.3m. Besides, we report the performance of orientation and translation errors for each scene, and compares them with [12], as can be seen in Table 3. The work of [12] is similar to our work, e.g., we both utilize the CNN features to select the most similar candidate frame, and use the matched feature points of the queue frame and candidate frame to compute the pose of queue frame. The main difference is that the method presented in [12] searches the whole database using CNN features to find the most similar candidate frame, and utilizes the matched keypoints of ORB features between the queue frame and candidate frame to estimate the pose, whereas our method proposes a coarse to fine searching mechanism which is more efficient. We first utilize the object information to exclude the unrelated frames in the database, then the CNN features are used to select the most similar candidate frame from the remaining frames in the database. At last, the keypoints of SURF features between the objects in the queue frame and candidate frame are matched to estimate the pose of queue frame.

For ICL-NUIM dataset, the mean error of orientation and translation of our proposed method is 1.7307° and 0.0628m, and the median error is 0.3327° and 0.0118m. For TUM RGB-D dataset, the mean error is 2.9210° and 0.1581m, and the median error is 0.6641° and 0.0291m. Our method can achieve more than 90% average localization success rate for most scenarios of ICL-NUIM dataset and TUM RGB-D dataset. For the officeRoom dataset, the number of objects is too few to have enough keypoints of recognized objects for pose estimation, which can be solved by using the keypoints of whole image for matching. Although the median error of orientation estimated by our method is higher than the work in [12], both the mean errors of orientation and translation of our proposed method are lower. The possible reasons are following: first, the SURF keypoints in general are more robust than ORB keypoints for illumination and scale change, so the mean errors of our method are lower. Second, our method only uses the keypoints of objects to accelerate localization speed. For the scenarios that do not have enough objects, we can not extract enough keypoints for matching, which can explain that the median error of our algorithm is higher.

We compare the proposed localization method with three CNN-based state of the art methods: (i) PoseNet [13]: a CNN model that directly regresses the camera pose with RGB input. (ii) 4D PoseNet [14]: a modified version of PoseNet with RGB-D input. (iii) CNN+LSTM [14]: it uses the PoseNet as the pose estimator and the LSTM as a temporal filter to estimate pose. Table 4 summarizes the mean errors of these methods on ICL-NUIM dataset. We can see that our method can achieve better accuracy both in orientation and translation compared to others.

We implement the proposed localization method on a computer with Intel Xeon E5-1650 v3 CPU @ 3.50GHz

TABLE 3. Localization performance for different scenarios.

Dataset	Scenario	Ours			Chen et al. [12]	
		The Median Error	The Mean Error	Success Rate	The Median Error	The Mean Error
ICL-NUIM	living room 'of kt0'	0.0096m 0.3265°	0.0607m 2.2212°	91.70%	-	-
	living room 'of kt1'	0.0103m 0.2780°	0.0396m 0.9997°	96.00%	-	-
	living room 'of kt2'	0.0097m 0.2612°	0.0395m 0.9361°	99.87%	-	-
	living room 'of kt3'	0.0088m 0.3142°	0.0317m 1.3179°	95.75%	-	-
	office room 'of kt0'	0.0199m 0.4123°	0.1549m 3.1138°	89.77%	-	-
	office room 'of kt1'	0.0091m 0.3193°	0.0462m 2.0639°	77.94%	-	-
	office room 'of kt2'	0.0113m 0.2948°	0.0768m 1.5928°	94.00%	-	-
	office room 'of kt3'	0.0152m 0.4550°	0.0525m 1.5997°	86.96%	-	-
	living room average	0.0096m 0.2950°	0.0429m 1.3687°	95.83%	0.05m 0.02°	0.36m 4.36°
	office room average	0.0139m 0.3704°	0.0826m 2.0926°	87.17%	0.07m 0.01°	0.31m 2.47°
	all images	0.0118m 0.3327°	0.0628m 1.7307°	91.50%	0.06m 0.01°	0.34m 3.43°
TUM RGB-D	freiburg1_plant	0.0179m 0.6079°	0.0678m 2.3161°	93.29%	0.12m 0.01°	0.38m 3.37°
	freiburg1_room	0.0141m 0.6978°	0.0768m 3.4290°	92.43%	0.17m 0.54°	0.43m 4.82°
	freiburg2_360_hemisphere	0.0797m 0.8259°	0.4790m 5.4154°	89.65%	0.05m 0.16°	0.38m 6.55°
	freiburg2_flowerbouquet	0.0138m 0.6443°	0.0295m 1.6028°	96.12%	0.07m 0.12°	0.15m 5.32°
	freiburg2_pioneer_slam3	0.0344m 0.6426°	0.2653m 3.7028°	85.76%	0.13m 0.13°	0.34m 8.80°
	freiburg3_long_office_household	0.0148m 0.5660°	0.0302m 1.0600°	99.67%	0.15m 0.21°	0.36m 3.00°
	all images	0.0291m 0.6641°	0.1581m 2.9210°	92.82%	0.10m 0.16°	0.32m 5.58°

TABLE 4. Comparison of mean errors for ICL-NUIM dataset.

Method	Living Room	Office Room
PoseNet	0.60m 3.64°	0.46m 2.97°
4D PoseNet	0.58m 3.40°	0.44m 2.81°
CNN+LSTM	0.54m 3.21°	0.41m 2.66°
ours	0.04m 1.37°	0.08m 2.09°

using a NVIDIA TITAN XP GPU. Based on the coarse to fine mechanism for excluding most images in database, the proposed method takes 296.5*ms* to find the most similar candidate frame and 277.9*ms* to estimate the final pose without optimization.

IV. CONCLUSION

This paper proposes a deep learning based visual semantic database construction method, and its application for visual localization with a coarse to fine mechanism. The visual semantic database includes the CNN features of image, the objects information recognized by Mask R-CNN and the SURF features of objects. To improve the algorithm efficiency, we use a coarse to fine mechanism to find out the candidate frames, i.e., a coarse searching step using object class and a fine searching step using the CNN features of image. Finally, a high accuracy localization result can be derived using the novel BA algorithm. We demonstrate our idea on the public datasets of ICL-NUIM and TUM RGB-D, which show improved performance considering the accuracy and efficient of the localization algorithm. Furthermore, the semantic database is also possible shared with other robots on a cloud platform, and can be further used for robot task reasoning, loop close detection, human-robot interaction and so on. In future, we would like to improve the SLAM system performance using image assessment and enhancement techniques to handle poor image quality problems due to camera shake or compression [28]–[30].

REFERENCES

[1] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 55–81, 2015.

[2] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[3] R. F. Salas-Moreno, "Dense semantic SLAM," Ph.D. dissertation, Dept. Comput., Imperial College London, London, U.K., 2014.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[5] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[7] Y. Zhao, G. Liu, and G. Tian, "A survey of visual SLAM based on deep learning," *Robot*, vol. 39, no. 6, pp. 889–896, 2017.

[8] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May/Jun. 2017, pp. 4628–4635.

[9] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-time dense SLAM and light source estimation," *Int. J. Robot. Res.*, vol. 35, no. 14, pp. 1697–1716, 2016.

[10] X. Li and R. Belaroussi. (2016). "Semi-dense 3D semantic mapping from monocular SLAM." [Online]. Available: https://arxiv.org/abs/1611.04144

[11] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 834–849.

[12] Y. Chen, R. Chen, M. Liu, A. Xiao, D. Wu, and S. Zhao, "Indoor visual positioning aided by CNN-based image retrieval: Training-free, 3D modeling-free," *Sensors*, vol. 18, no. 8, p. 2692, 2018.

[13] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2015, pp. 2938–2946.

[14] F. Guo, Y. He, and L. Guan, "RGB-D camera pose estimation using deep neural network," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2017, pp. 408–412.

[15] H. Taira *et al.*, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proc. CVPR*, Jun. 2018, pp. 7199–7209.

[16] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep./Oct. 2015, pp. 4297–4304.

[17] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

[18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[19] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[21] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[22] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Proc. Int. Workshop Vis. Algorithms*. Berlin, Germany: Springer, 1999, pp. 298–372.

[23] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Hong Kong, May/Jun. 2014, pp. 1524–1531.

[24] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2012, pp. 573–580.

[25] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

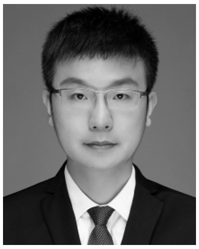
[26] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 936–944.

[28] K. Bahrami and A. C. Kot, "Efficient image sharpness assessment based on content aware total variation," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1568–1578, Aug. 2016.

[29] Q. Wu *et al.*, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[30] K. Gu, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "Automatic contrast enhancement technology with saliency preservation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 9, pp. 1480–1494, Sep. 2015.



WEI ZHANG received the B.Sc. degree from the College of Information and Control Engineering, China University of Petroleum, Qingdao, Shandong, China, in 2016. He is currently pursuing the master's degree with the School of Control Science and Engineering, Shandong University, Jinan, Shandong. His research interests include simultaneous localization and mapping and deep learning.

GUOLIANG LIU (S'11–M'12) received the B.Sc. degree in physics from Shandong Normal University, Jinan, China, in 2005, the M.Eng. degree in control science and engineering from the National University of Defense Technology, Changsha, China, in 2008, and the Ph.D. degree in Computer Science from the University of Göttingen, Göttingen, Germany, in 2012. He was a Staff Researcher with Lenovo Group, from 2012 to 2014. He is currently an Associate Professor of control science and engineering with the School of Control Science and Engineering, Shandong University, Jinan. His research interests include distributed camera networks, information fusion, human action recognition, human–robot safe interaction, and autonomous robot.

GUOHUI TIAN received the M.S. degree in industry automation from Shandong University, Jinan, China, in 1993, and the Ph.D. degree in automatic control theory and application from Northeastern University, Shenyang, China, in 1997. He was a Postdoctoral Researcher with the Engineering Department, Tokyo University, from 2003 to 2005. He is currently a Professor with the School of Control Science and Engineering, Shandong University. His research mainly focuses on service robots and smart space.