Proceedings of The 2021 IEEE International
Conference on Real-time Computing and Robotics
July 15-19, 2021, Xining, China

A Hierarchical Semantic Map for Large-scale Outdoor Environment

Di Zhang, De Xu

***Abstract*—We make full use of the pre-known knowledge of environment, and establish a hierarchical semantic map offline for large-scale outdoor environment. The map contains semantic information which is more stable than the commonly used feature points. And the description and recognition methods of locations based on semantic information are similar to human habits. Experiment results show that the all parts of the map working together can achieve path planning, location recognition and relative pose estimation to complete the robot navigation task.**

I. Introduction

The common map models of environment can generally be divided into metric maps, topological maps and semantic maps. Using metric maps can construct a map precisely when dealing with relatively small environment, and can estimate the relative pose of the robot's trajectory in the map. However, when dealing with large-scale environment, global metric map is likely to cause problems such as positioning failure due to poor global consistency, and its robustness cannot be satisfactorily guaranteed. Topological map emphasizes the relationship between nodes and edges, ignoring most of the scale information in the map, and it is difficult to determine the robot's position accurately in environment. When performing navigation and positioning tasks in a large-scale environment, the semantic information of objects in the environment is what humans rely on when they search the path in environment. Semantic information is also a valid clue that can improve robustness.

In order to jointly adopt the advantages of both the metric map and the topological map, Zhou *et al.* [1] proposed a novel grid-topological map, using pre-known knowledge of the environment, and developed an accurate and efficient indoor pathfinding scheme based on building information modeling data. A sparse semantic map building method and a laser relocalization strategy were proposed in [2]. After initial classification and reclassification on 3D point clouds, the outdoor environments were divided into scene nodes and road nodes by scene understanding. According to the generating topological relations between the scene nodes and the road nodes, the semantic map of the outdoor environment was established. Then the map was simplified so that the positioning accuracy could be improved.

In order to make full use of the pre-known knowledge of environment and the advantages of various maps, we propose a novel hierarchical semantic outdoor map. We use pre-known satellite picture to generate topological maps and a simple global metric map, then we can use them for path planning. We extract HOG [3] features of the surroundings near the nodes in real world as for semantic features to construct semantic map. The semantic information near each node is used for node recognition according to the semantic map. After confirming which node is the robot located, the local metric map of the node is used to estimate the relative pose and determine accurate position of the robot in the map.

The rest of this article is organized as follows. Section II discusses related works. Section III specifically introduces our approach. Section IV shows the experiment results, and, finally, Section V concludes this article.

II. Related Works

There have been a lot of discussions about the mapping and localization of large-scale outdoor environment. Metric maps are mostly used. In many works, localization or navigation task is based on the laser [4, 5]. Reference [4] proposed a fusion method to combine the RTK-GPS (RTK: Real-time kinematic, GPS: Global positioning system) with the laser based simultaneous localization and mapping (SLAM) method to design a localization mechanism to make a four-wheel-independent-drive mobile robot run free to the outdoor environment. Reference [5] improved a high-precision lightweight laser SLAM method for outdoor large-scale scenes. Camera is a more inexpensive choice than laser, providing texture-rich information about scene at different distance and is widely used in large-scale outdoor environment [6-8]. LSD-SLAM (LSD: Large-scale direct monocular) [6] can not only locally track the motion of the camera, but also allow to build consistent, large-scale maps of the environment.

Topological maps are also used to build maps of outdoor large-scale scenes [9-13]. Reference [9] presented a real-time hierarchical (topological/metric) SLAM system based on the fusion of stereovision and GPS in order to realize the autonomous vehicle outdoor navigation in large-scale environment, keeping both the local consistency and the global consistency. Reference [10] presented a semantic map management approach for various environments by triggering multiple maps with different SLAM configurations, combining laser, visual, IMU (IMU: Inertial measurement unit) sensors together, to realize the navigation both in indoor and outdoor environment.

Location recognition based on semantic information is similar to the way of human. In recent years, semantic maps are paid more attention than before [2, 14-20]. Reference [2] built a sparse semantic map from three-dimensional (3D) point clouds through scene understanding process by classifi-

*This work was supported in part by the Science and Technology Program of Beijing Municipal Science and Technology Commission (Z191100008019004).

All authors are with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: de.xu@ia.ac.cn).

2021 IEEE International Conference on Real-time Computing and Robotics (RCAR) | 978-1-6654-3678-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/RCAR52367.2021.9517476

cation. X-View [15] leveraged semantic graph descriptor matching for global localization, enabling localization under drastically different view-points. Reference [20] proposed a method with a new multi-task semantic and depth prediction model and a superpixel-based refinement for monocular semantic mapping.

III. Approach

In this section, we propose a novel hierarchical semantic outdoor map which consists of topological map, simple global metric map, semantic map, and local metric map, termed topological- semantic map.

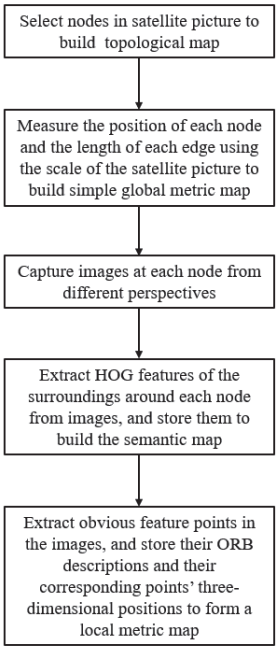


Figure 1. Overall framework of the topological-semantic map

A. Map Building

In topological-semantic map, the topological map and simple global metric map are firstly constructed from satellite picture. After extracting the roads in the satellite picture with pre-known knowledge, the outdoor space is modeled as a topography with nodes and edges selected manually. A node usually denotes an important spatial position such as intersection, turning or building entrance. Edges, representing roads, connected each node. The topological map is represented by an undirected graph with edges and nodes. The accurate location of each node can be obtained by satellite picture and its scale, then the simple global metric map can be established. Robots can operate path planning according to the topological map and the simple global metric map. People often use the nearby surroundings of a place as semantic information when they describe it. A unique place can be identified by comparing the surroundings' information. In the proposed method, surroundings' information of each node can be encoded as semantic information and can be used in node recognition. Landmarks around each node, for example, trees and buildings, are manually selected from images collected in the real world and transformed to image features to represent semantic objects. We choose HOG features as semantic information and use them to describe and distinguish each node, for different node has its different surroundings' appearances. These HOG features of landmarks are stored in the map to form a semantic map. The examples of HOG features selection are shown in Figure 2. By searching semantic objects in the images collected by the robot from the real world and comparing them with semantic objects around each node, the node which the robot is located at can be determined. The obvious feature points in the images obtained around the node are matched with the corresponding points in the real world. The ORB description of each feature points extracted in the frame and its corresponding point's three-dimensional position are stored in the map to form a local metric map. When the robot is close to the nodes, feature points extracted in the frame can be matched to feature points stored in the map, and can be used to estimate the relative pose then the accurate position of the robot in the map can be obtained. These maps work together to realize path planning and positioning for robots in large-scale outdoor environment. The overall process of map building is shown in Figure 1.

Figure 2. The examples of HOG features selection

B. Path Planning

Different from metric maps, topological maps can not achieve as high accuracy as metric maps do. However, the way of path planning with topological map is similar to that of human beings. Currently, the topological map has been widely used by various navigation tasks. The Dijkstra's algorithm can be used for effective path planning based on topological maps to get the shortest path between two nodes in the map. But when we select the starting point and the ending point randomly in the map, the points we selected are not the nodes in the adjacency matrix $\boldsymbol{G}$ of the topological map. Before performing the Dijkstra's algorithm, we find the closest points of the starting point and the ending point which are on the edges to generate new nodes and new edges, and add the new nodes and new edges to the adjacency matrix $\boldsymbol{G}$.

The algorithm works as follows:

Algorithm I. Path Planning Process.

Input: starting node $\boldsymbol{s}$, ending node $\boldsymbol{e}$, the adjacency matrix $\boldsymbol{G}$ of the topological map, the number nodes n, the number of edges l.

Output: nodes that the shortest path passes ***path*** and the shortest length l_{shortest}.

1 $\boldsymbol{G}$ = initialize($\boldsymbol{G}$, $\boldsymbol{s}$, $\boldsymbol{e}$), $n = n+2$, $l = l+2$

2 $\boldsymbol{S}$ = [ ], ***dist*** =[∞], ***prev*** = [ ]

3 **for** i = 0,1,…,n-1

4 dist[i] = $length_{s,\text{I}}$, $\boldsymbol{S}$ [i] = 0

5 **if** dist[i] == ∞

6 prev[i] = 0

7 **else**

```
        prev[i] = s;
    end if
end for
dist[s] = 0, s[s] = 1
for i = 1,2,…,n-1
  tmp = ∞, u = v
  for j = 0,1,…,n-1
    if s[j]=0 && dist[j]<tmp
      u = j, tmp = dist[j]
    end if
  s[u] = 1
  for j = 0,1,…,n-1
    if s[j]=0 && length_{u,j}<∞
      newdist = dist[u] + length_{u,j}
      if newdist < dist[j]
        dist[j] = newdist, prev[j] = u
      end if
    end if
  end for
que[n], path[], tot = 1, que[tot] = e , tot++, tmp = prev[e]
while tmp != s
que[tot] = tmp, tot++, tmp = prev[tmp]
end while
que[tot] = e
for i = tot,tot-1,…,1
  path.push_back(que[i])
end for
return l_shortest =dist[n], path
```

C. Node Recognition

In the real world, the appearances of surroundings change greatly when observing from different positions in different directions. The surroundings around a node can be a unique semantic description of the node, and object features are more stable than point features. In order to distinguish each node effectively, we have counted which types of surrounding objects around each node, and converted the statistical results into a description $d_k=(d_1,d_2,\ldots,d_N)^T$, k representing the kth node, and $d_i = 1$ or 0 representing the ith object is observed or not. Since object classification based on deep learning requires a large amount of data, it is very inconvenient and expensive to use it. Outdoor objects are trees and buildings in majority, and their contours are easy to distinct. So we use HOG features to identify the objects. We select the representative HOG feature for each object as a positive sample, and classify them by comparing the cosine distance between the test sample and the positive sample. The cosine distance can be calculated as:

$$\cos(m,n)=\frac{|m\cdot n|}{\|m\|\cdot\|n\|} \tag{1}$$

Where m is the test sample, and n is the positive sample.

After selecting HOG features for all the objects around the nodes in the map, we have a set of indexes for all the features, $O = \{O_i\}$, and which node is the object located in, $L = \{L_i\}$, i=1,2,…,N, for N is the total number of the objects.

The number of HOG features of each node is sorted according to the angle range which the object can be seen from large to small, and we give a set $P_k=\{O_p^k\}$, where $O_p^k \in \{O_i\}$. For each object, we record the other objects which can be seen with it in the same perspective, given a set of objects, $Q_i = \{O_q\}$, $O_q \in \{O_i\}$.

Given three images in different directions, the process of node recognition is explained below:

Algorithm II. Node Recognition Process.

Input: the total objects set $\{O_i\}$, $\{L_i\}$, objects set of each node $P_k=\{O_p^k\}$, nodes number K, imges set $\{C_1,C_2,C_3\}$, threshold $\{th_i\}$, $\{Q_i\}$

Output: the node which the robot located, k

```
for i = 1,2,…,N
  hog_num[O_i] = 0, complete[O_i]=0
end for
for p = 1,2,…,N
  for k = 1,2,…,K
    for n = 1,2,3
      cos[n][k] = max(cos(fv[n],fv[O_p^k]))
    end for
    complete[O_p^k] = 1
  end for
  for k = 1,2,…,K
    for n = 1,2,3
      if cos[n][k] > th[O_p^k]
        hog_num[k] = hog_num[k]+1, d_k'[O_p^k]=1
        if Q_Opk != ∅
          for j=1,2,…,size of Q_Opk
            pre[].pushback(Q_Opk[j])
          end for
        end if
      end if
    end for
  end for
  if pre!= ∅
    for j=1,2,…,size of pre
```

```
        if complete[O_j] = 0
          for n = 1,2,3
            complete[O_j] = 1
            if max(cos(fv[n],fv[O_j])) > th[O_j]
          hog_num[L[O_j]] = hog_num[L[O_j]]+1, d L[Oj]·[O_j]=1
            end if
          end for
        end if
      end for
    end if
    for k = 1,2,…,K
      if hog_num[k] > 2
        if cos(d_k, d_k') > threshold
          break
        end if
      end if
    end for
  end for
return k
```

D. Relative Pose Estimation

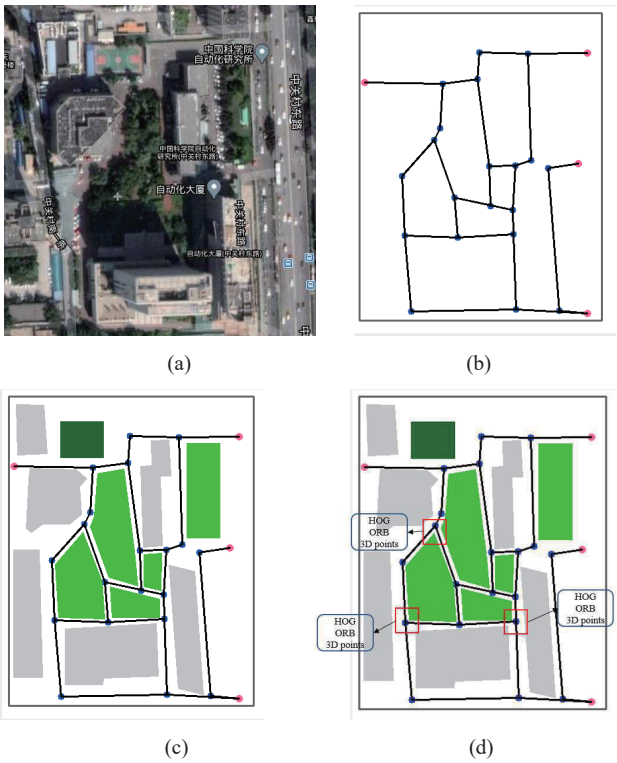
Perspective-n-point (PnP) method measures the position of camera based on the pre-known knowledge of the objects. In the PnP method, by using n given points whose positions are known in reference frame, the pose of reference frame relative to the camera frame can be determined. For example, 4 corners of a rectangle were used to computed the robot's pose with PnP method in [21]. After establishing a coordinate near each node in the map, we choose obvious points on the reference objects near the origin of node coordinate system, store their ORB features and positions in the coordinate system, then we utilize PnP method on these points after match them with the points in camera frame, so that the relative pose of robot to the coordinate system can be estimated easily.

IV. Experiment

The proposed approach has been implemented in the area larger than 1000 square meters on a computer with 3.6 GHz CPU, 8 GB RAM, and 64-bit operating system. In this section, we show the results to confirm the feasibility of the proposed method.

A. Map Building

Applying the approach in section III, we built the topological- semantic map of an area larger than 1000 square meters. The process of map building is shown in Figure 3. The satellite picture is shown in Figure 3(a), and its corresponding simple global metric map is shown in Figure 3(b). After adding semantic object to the map, the semantic map is shown in Figure 3(c). The gray areas represent buildings and green areas represent plants. Figure 3(d) shows the sketch map of the semantic map and the local metric map. HOG features of the surroundings of each node were stored to build the semantic map. As for the obvious feature points in the images captured near each node, their ORB descriptions and their corresponding points' three-dimensional positions were stored for the local metric map.

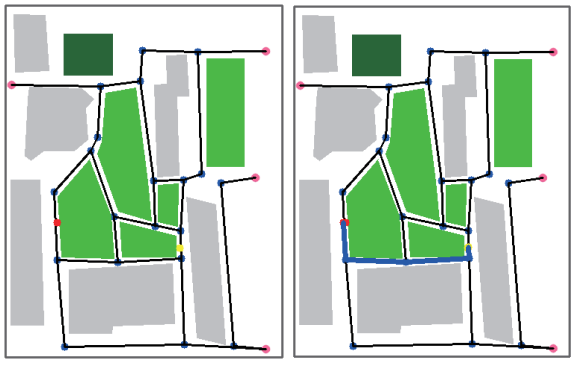


(a) (b) (c) (d)

Figure 3. The process of map building, (a) Satellite picture (b) simple global metric map (c) semantic map (d) the sketch map of the semantic map and the local metric map

B. Path Planning

We randomly selected the starting point and the ending point in the interface, and used the method mentioned in section III to plan the path. The result is shown in Figure 4. The red point represents the start point, and the yellow point represents the end point, as shown in Figure 4(a). The planned path is shown in Figure 4(b), which is the shortest route marked with a bold blue line in the map.



(a) (b)

Figure 4. The result of path planning, (a) the start and target points, (b) the planned path

It can be seen that the proposed method can effectively find the shortest path.

C. Node Recognition

For HOG feature, in our experiments, we used 64×64 sized image, 16×16 sized cell, 32×32 sized block, and got a 324-dimensional vector for each 64×64 sized image. There were three images in each set, and the difference of angle when one of the images was taken between the two others was at least over 45 degrees, as shown in Figure 5.

Figure 5. A set of images of node1

We selected three nodes to verify the robustness of the method, and each node contained 30 sets of images. In the robustness test, the different images including the images did not belong to the node were selected to form a testing set of images. We selected 15 testing sets of other images. The test results show that these sets of other images are not recognized as the node. The sets of images belonging to the node are correctly recognized. The results of node recognition are shown in the following table. Our method can accurately distinguish each node.

TABLE I. NODE RECOGNITION RESULTS

Node	Correct rate
1	100%
2	100%
3	100%

D. Relative Pose Estimation

For relative pose estimation, we tested six frames applying the PnP method to get the position in XOY plane relative to the coordinate of the node. The reference objects were the pillars as shown in Figure 6. The origin of the reference frame was set to the place at the left pillar. The robot moved to different locations relative the pillars. Its relative positions in the reference frame were manually measured with a rule, which were taken as the ground truth. The positions of the top and bottom of pillars were known as the pre-knowledge. They were used in the PnP method. The pillars' images were captured and the robot's positions in the reference frame were calculated with the PnP method. The results are showed in TABLE II. All the relative errors to the depth are less than 5%.

Figure 6. The reference object near the node

TABLE II. RELATIVE POSITION ESTIMATION RESULTS

Ground truth(mm)		Measured position (mm)		Error(mm)	
x	y	x	y	Δx	Δy
3225	4844	3124.6	4954.8	-100.4	110.8
2465	3992	2402.2	4091.9	-62.8	99.9
1959	3494	1982.2	3586.0	23.2	92.0
1950	3206	2026.7	3257.9	76.6	51.9
1940	3408	2025.6	3445.0	85.6	37.0
1914	3591	1996.8	3636.7	82.8	45.7

V. CONCLUSION

We propose a mapping method for large-scale outdoor environment, which is convenient and quick. The experiments show that sub-maps in our map can work together to complete tasks such as path planning, positioning, and relative pose estimation successfully.

REFERENCES

[1] X. Zhou, Q. Xie, M. Guo, J. Zhao and J. Wang, "Accurate and Efficient Indoor Pathfinding Based on Building Information Modeling Data," *IEEE Trans. Industrial Informatics*, vol. 16, no. 12, pp. 7459-7468, Dec. 2020.

[2] F. Yan, J. Wang, G. He, H. Chang, & Y. Zhuang, "Sparse semantic map building and relocalization for UGV using 3D point clouds in outdoor environments," *J. Neurocomputing*, 2020

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, San Diego, CA, USA, pp. 886-893 vol. 1.

[4] Y. Deng, Y. Shan, Z. Gong and L. Chen, "Large-Scale Navigation Method for Autonomous Mobile Robot Based on Fusion of GPS and Lidar SLAM," *2018 Chinese Automation Congress (CAC)*, Xi'an, China, 2018, pp. 3145-3148.

[5] C. Pang, Y. Tan, S. Li, Y. Li, B. Ji and R. Song, "Low-cost and High-accuracy LIDAR SLAM for Large Outdoor Scenarios," *2019 IEEE Int. Conf. Real-time Computing and Robotics (RCAR)*, Irkutsk, Russia, pp. 868-873.

[6] J. Engel, T. Schöps and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM", *European Conf. Computer Vision*, 2014.

[7] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg and V. Lepetit, "Instant Outdoor Localization and SLAM Initialization from 2.5D Maps," *IEEE Trans. Visualization and Computer Graphics*, vol. 21, no. 11, pp. 1309-1318, 15 Nov. 2015.

[8] A. Armagan, M. Hirzer, P. M. Roth and V. Lepetit, "Learning to Align Semantic Segmentation and 2.5D Maps for Geolocalization," *2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, pp. 4590-4597.

[9] D. Schleicher, L. M. Bergasa, M. Ocana, R. Barea and M. E. Lopez, "Real-Time Hierarchical Outdoor SLAM Based on Stereovision and GPS Fusion," *IEEE Trans. Intelligent Transportation Systems*, vol. 10, no. 3, pp. 440-452, Sept. 2009.

[10] S. F. G. Ehlers, M. Stuede, K. Nuelle and T. Ortmaier, "Map Management Approach for SLAM in Large-Scale Indoor and Outdoor Areas," *2020 IEEE Int. Conf. Robotics and Automation (ICRA)*, Paris, France, pp. 9652-9658.

[11] N. Islam, K. Haseeb, A. Almogren, I. U. Din, M. Guizani, & A. Altameem, "A framework for topological based map building: A solution to autonomous robot navigation in smart cities," *J. Future Generation Computer Systems*, 2019.

[12] V. Balaska, L. Bampis, M. Boudourides, & A. Gasteratos, "Unsupervised semantic clustering and localization for mobile robotics tasks," *J. Robotics and Autonomous Systems*, 103567, 2020.

[13] C. Boucher, & J.-C. Noyer, "Automatic Detection of Topological Changes for Digital Road Map Updating," *IEEE Trans. Instrumentation and Measurement*, 61(11), 3094–3102,2012.

[14] C. Zhang, Z. Liu, G. Liu and D. Huang, "Large-Scale 3D Semantic Mapping Using Monocular Vision," *2019 IEEE 4th Int. Conf. Image, Vision and Computing*, Xiamen, China, pp. 71-76.

[15] A. Gawel, C. D. Don, R. Siegwart, J. Nieto and C. Cadena, "X-View: Graph-Based Semantic Multi-View Localization," *IEEE J. Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687-1694, July 2018.
[16] KN. Lianos, J.L. Schönberger, M. Pollefeys, T. Sattler, "VSO: Visual Semantic Odometry," *European Conf. Computer Vision*, 2018.
[17] E. Stenborg, C. Toft and L. Hammarstrand, "Long-Term Visual Localization Using Semantically Segmented Images," *2018 IEEE Int. Conf. Robotics and Automation*, Brisbane, QLD, pp. 6484-6490.
[18] K. Doherty, D. Fourie and J. Leonard, "Multimodal Semantic SLAM with Probabilistic Data Association," *2019 Int. Conf. Robotics and Automation*, Montreal, QC, Canada, pp. 2419-2425.
[19] F. Bernuy and J. Ruiz Del Solar, "Semantic Mapping of Large-Scale Outdoor Scenes for Autonomous Off-Road Driving," *2015 IEEE Int. Conf. Computer Vision Workshop*, Santiago, pp. 124-130.
[20] Y. Bai, L. Fan, Z. Pan and L. Chen, "Monocular Outdoor Semantic Mapping with a Multi-task Network," *2019 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Macau, China, pp. 1992-1997
[21] G. Chen, D. Xu, P. Yang, "High precision pose measurement for humanoid robot based on PnP and OI algorithms," *2010 IEEE Int. Conf. Robotics and Biomimetics*, Tianjin, China, 2010, pp. 620-624.