AN INDOOR SCENE LOCALIZATION METHOD USING GRAPHICAL SUMMARY OF MULTI-VIEW RGB-D IMAGES

Preeti Meena, Himanshu Kumar, Sandeep Yadav

Indian Institute of Technology Jodhpur, Rajasthan, India

1. STUDY OF SIMILAR OBJECTS IN MULTI-VIEW IMAGES

To decide an optimal threshold for fusing two nodes/objects of two different views, we have performed a study that analyzes the similarity between objects belonging to different views. For this study, we have computed the Chamfer distance [1] between 268 object pairs from the multi-view scene; among them, 132 represents the pair of similar objects while 136 represents the pair of objects that are different from each other. Similar objects are those objects that have the same characteristics and should be fused to generate a unique multi-view graphical summary. Figure 1 illustrates the histograms of Chamfer distance [1] for the pair of similar/matched objects (data 1) and unmatched objects (data 2) among multiple views. We observed from the figure that the optimal threshold τ obtained by using maximum likelihood estimation (MLE) [2] is 1.89. Thus, this heuristically shows that the assumption about the threshold value used in subsec. 4.1.(B), based on the obtained distance, is appropriate for node fusion.

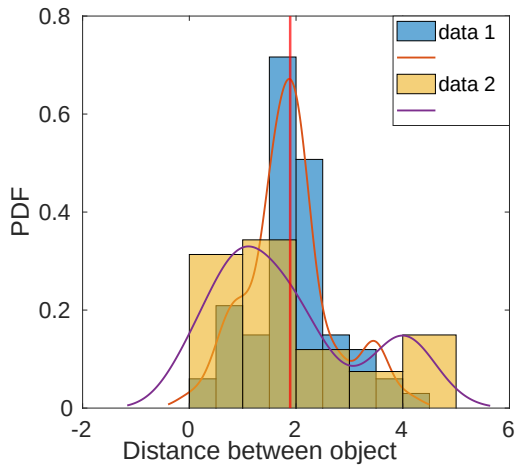


Fig. 1. Threshold estimation for node fusion.

2. MULTI-VIEW GRAPH DATASET STRUCTURE

We have utilized scenes from the SUNRGB-D [3] dataset to create a multi-view graph database of 10 buildings. A total of 224 RGB-D images are included for 10 buildings. Each building consists of several rooms that define different indoor scenes, e.g. kitchen, bedroom, etc., and each room/scene includes a set of multi-view RGB-D images. The structure of the created multi-view graph database for a building is shown in Fig. 2, in which the multi-view graph is a compact version (summary) of a scene that includes adequate information extracted from all views for representing complete scene information. A few examples of a generated summary are shown in Fig. 3. From the figure, we observe that a scene consisting of more than one view is successfully represented by a single graphical summary. These generated multi-view graphical summaries are further utilized to perform indoor scene localization. The pseudocode for the proposed graphical summary-based scene localization is given in 1.

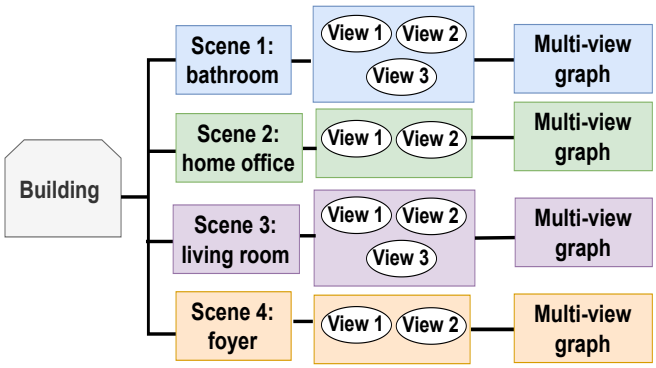


Fig. 2. Multi-view graph dataset structure for one building consists of four scenes.

3. RESULTS

This section presents the results obtained using the proposed graphical summary-based scene localization and state-of-the-art methods including [5], [6], [7], [8], [9], and [10] on few sample query images. Figure 4 shows the predicted location for 12 query images belonging to 3 buildings. We observe that the proposed method outperforms others.

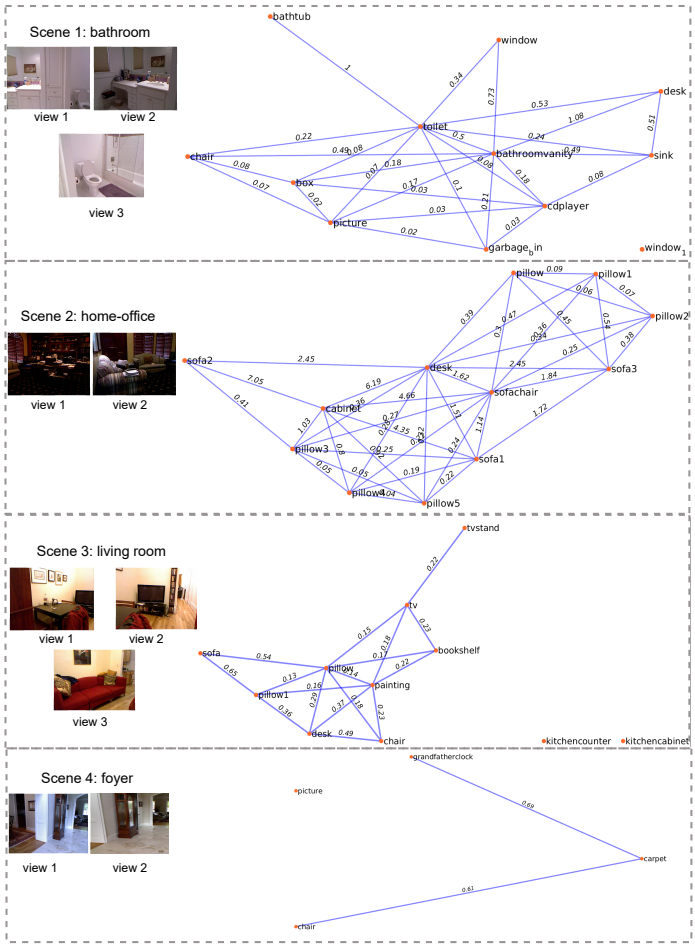


Fig. 3. Generated multi-view graphical summary (right) for four scenes (left) of a building.

Algorithm 1: Indoor scene localization using graphical summary of multi-view RGB-D images.

Input: Input multi-view RGB-D image set I

Output: Scene location $\mathcal{S}^g$

Step1: Single-view graph construction, $G^{v_i} = (\mathcal{N}^{v_i}, E^{v_i})$;

(a). Extract salient objects $\{\mathcal{O}_i\}_{i=1}^{n} \in \mathcal{N}^{v_i}$ via $\breve{S}\{O_i\}$ [4];

(b). Compute adjacency matrix such that

$$E^{v_i}(i,j) = \begin{cases} 1; & \text{if } \min(l_i, l_j) > d_s(i,j) \\ 0; & \text{Otherwise} \end{cases} \quad (1)$$

Step2: Multi-view graph construction, $G = (\mathcal{N}, E, w)$ using an approach given in sec. 4.1.(B);

Step3: Scene localization for a query graph G^q;

(a). Candidate matching, $\hat{i} = \arg\max_i \ \mathcal{J}(G^q, G_i^{\mathcal{D}})$;

(b). Node alignment,

$$\mathcal{G}_{\hat{i}}^c \subseteq \mathcal{G}_{\hat{i}}^{\mathcal{D}} \text{ s.t.} A_{\mathcal{G}_{\hat{i}}^c}(k,j) = \begin{cases} 1 & \text{if } j = \arg\max_j \mathbf{d}^{\hat{i}}(f_k, \tilde{f}_j) \\ 0 & \text{Otherwise.} \end{cases};$$

(c). Scene graph matching, $\hat{i} = \arg\max_{\hat{i}} \ \mathcal{M}(G^q, G_{\hat{i}}^c)$ and $\mathcal{S}^g = \mathcal{S}_{\hat{i}} \in \mathcal{B}$

4. REFERENCES

[1] Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua, "Point-set distances for learning representations of 3d point clouds," in *ICCV*, 2021, pp. 10478–10487. 1

[2] In Jae Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003. 1

[3] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015, pp. 567–576. 1

[4] Preeti Meena, Himanshu Kumar, and Sandeep Yadav, "A volumetric saliency guided image summarization for rgb-d indoor scene classification," *arXiv preprint arXiv:2401.16227*, 2024. 2

[5] Meng-Jiun Chiou, Zhenguang Liu, Yifang Yin, An-An Liu, and Roger Zimmermann, "Zero-shot multi-view indoor localization via graph location networks," in *ACMMM*, 2020, pp. 3431–3440. 1, 3

[6] Dapeng Du, Limin Wang, Zhaoyang Li, and Gangshan Wu, "Cross-modal pyramid translation for rgb-d scene recognition," *IJCV*, vol. 129, no. 8, pp. 2309–2327, 2021. 1, 3

[7] Albert Mosella-Montoro and Javier Ruiz-Hidalgo, "2d–3d geometric fusion network using multi-neighbourhood graph convolution for rgb-d indoor scene classification," *Information Fusion*, vol. 76, pp. 46–54, 2021. 1, 3

[8] Ali Caglayan, Nevrez Imamoglu, Ahmet Burak Can, and Ryosuke Nakamura, "When cnns meet random rnns: Towards multi-level analysis for rgb-d object and scene recognition," *Computer Vision and Image Understanding*, vol. 217, pp. 103373, 2022. 1, 3

[9] Bo Miao, Liguang Zhou, Ajmal Saeed Mian, Tin Lun Lam, and Yangsheng Xu, "Object-to-scene: Learning to transfer object knowledge to indoor scene recognition," in *IROS*. IEEE, 2021, pp. 2069–2075. 1, 3

[10] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra, "Omnivore: A single model for many visual modalities," in *CVPR*, 2022, pp. 16102–16112. 1, 3

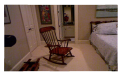
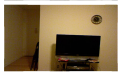
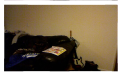
Query Images	Ground truth scene labels	Predicted scene labels [5]	[6]	[8]	[7]	[9]	[10]	Proposed
	B1S4	B1S1	B1S2	B1S4	B1S2	B1S4	B1S4	B1S4
	B2S3	B2S3	B2S5	B2S2	B2S3	B2S5	B2S3	B2S3
	B1S1	B1S1	B1S1	B1S1	B1S1	B1S1	B1S1	B1S1
	B1S5	B1S2	B1S2	B1S5	B1S2	B1S5	B1S2	B1S5
	B2S4	B2S1	B2S2	B2S2	B2S4	B2S1	B2S4	B2S4
	B2S4	B2S2	B2S2	B2S2	B2S2	B2S4	B2S4	B2S4
	B3S1	B3S1	B3S1	B3S1	B3S1	B3S1	B3S1	Not present
	B3S1	B3S1	B3S1	B3S1	B3S1	B3S1	B3S1	B3S1
	B2S5	B2S2	B2S3	B2S3	B2S5	B2S3	B2S3	B2S5
	B1S5	B1S2	B1S4	B1S5	B1S4	B1S2	B1S4	B1S5
	B2S1	B2S1	B2S3	B2S3	B2S1	B2S1	B2S1	B2S1
	Not present	B2S2	B2S2	B2S5	B2S5	B2S2	B2S5	Not present

Fig. 4. A few examples of scene localization. Top four rows: multi-view dataset for three buildings. Fifth to bottom rows: query images with ground truth scene label and predicted scene labels using methods in [5], [6], [8], [7], [9], and [10] and proposed method, (B=building, S=scene).