IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884183

MAKE OBJECT CONNECT: A POSE ESTIMATION NETWORK FOR UAV IMAGES OF THE OUTDOOR SCENE

Jingyi Cao[1, *], Yanan You[1], Le Xia[2], Jun Liu[1]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications,
Beijing, China
[2]Hunan Key Laboratory of Remote Sensing Monitoring of Ecological Environment in Dongting Lake
Area, Hunan Natural Resources Affairs Center, Changsha, China
*Corresponding author: caojingyi@bupt.edu.cn

ABSTRACT

As the basics of 3D vision, pose estimation with 2D images is of significance in 3D reconstruction, UAV positioning, and other fields. However, the related works focus on the natural images and pay less attention to the wide-coverage UAV remote sensing (RS) images. In fact, the relationship between objects in UAV images can benefit pose estimation. Therefore, aiming at the outdoor scene captured by the UAV monocular camera, a novel pose estimation network that emphasizes the association between objects is proposed. The multi-scale visual features extracted by the convolutional neural network (CNN) are manipulated by the object-agnostic segmentation model to indicate the existing space of all possible objects in the whole scene. The features of all possible objects are embedded into vectors, and then processed with a graph convolution network (GCN) for relationship analysis. Based on the known sparse point cloud and the optimized features of 2D images, the camera pose is regressed iteratively by 3D visual geometry. To verify the feasibility of the network, experiments are conducted on the Extended CMU Seasons and the simulation UAV dataset. Results prove that our network emphasizes more features on the small objects and obtains superior pose estimation results.

Index Terms— pose estimation, feature matching, unmanned aerial vehicle (UAV), 3D reconstruction

1. INTRODUCTION

Imaging the surface with Unmanned Aerial Vehicle (UAV) is an effective solution to investigating the environment and object morphology. The development of the computation resource makes 3D scene analysis with RS data possible. As basic of scene analysis, the 2D-2D pose estimation task resolves the relative pose transform of the camera through feature matching between 2D images. Nowadays, a lot of research has been done about the pose estimation of natural images. They mainly use conventional algorithms, like SIFT and SURF, to extract features and then estimate the motion relationship between monocular camera frames based on Epipolar Geometry [1]. However, they regard feature extraction and pose estimation as independent tasks. Recently, some studies have proved the ability of the neural network in pose estimation. They regress poses with an end-to-end CNN [2] or obtain camera pose after CNN-based feature matching [3]. However, these schemes disregard the 3D geometry theory. Especially, PixLoc [4] optimizes feature extraction under the guidance of 3D geometry, bridging the gap between CNN and 3D geometry.

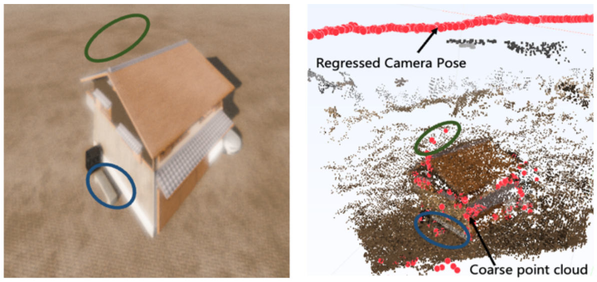

Fig.1 Phenomenon analysis. Left: image example. Right: 3D reconstruction result (red points: sparse reconstruction result from SFM, color point: dense reconstruction result from PMVS).

In fact, few studies related to pose estimation focus on RS data, especially UAV images. In our experiment, it is found when reconstructing the 3D scene with 2D images, the prejudice on large objects always happens, i.e., the algorithms tend to observe large objects but ignore small objects (blue ellipse in Fig.1). Feature point drift also occurs (red ellipse), which causes deviation in pose estimation. The phenomenon is ubiquitous in both natural and RS image analysis.

UAV RS images cover a wide range and with complex object distribution. Considering that individuals in RS images are relatively independent, each object provides important information for the analysis of the whole scene. Meanwhile, their relationship is also noteworthy. It is crucial to emphasize the existence of each object, balance the cognitive bias of the multi-scale targets during feature extraction, and then optimize feature extraction according to the correlation between targets. Therefore, aiming at the outdoor scene captured by UAV camera, an end-to-end pose estimation network based on object connection is proposed. With the generated sparse point cloud and 2D images, the CNN is adopted to extract image features, the object-agnostic segmentation is used for perceiving possible objects. The affinity between the possible objects and the corresponding point cloud is analyzed through GCN. Then, the relative poses of the cameras are analyzed based on geometric relation and optimized with bundle adjustment.

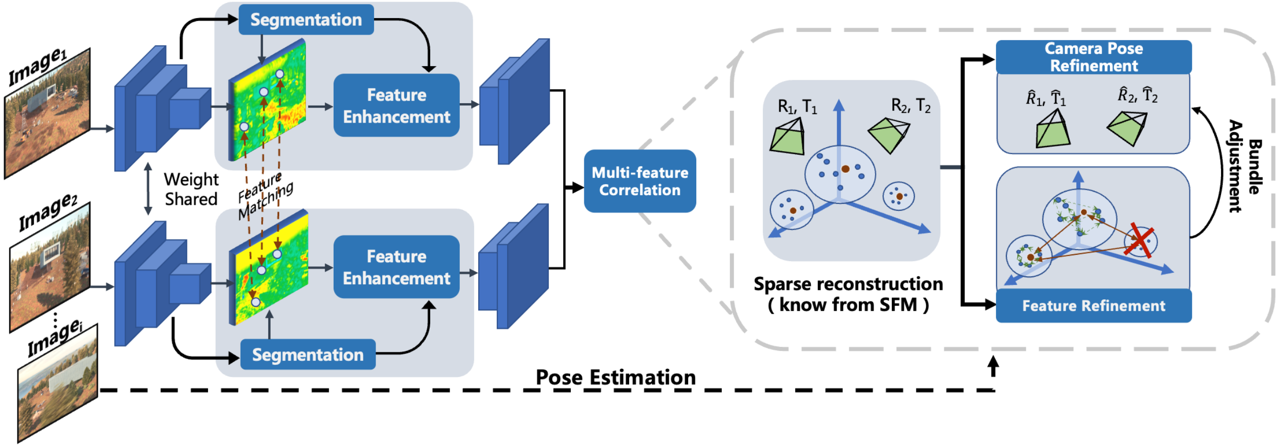

Fig. 2 The overall pose estimation network.

2. NETWORK ARCHITECTURE

In this section, we introduce the proposed network in detail. The overall network is shown in Fig.2, which is designed based on the encoder-decoder structure. It mainly contains two parts. One is the feature extraction and enhancement module, and the other is the multi-feature correlation module.

2.1. Feature Extraction and Enhancement Module

VGG16 is used as the basic feature extraction network. Inspired by the feature pyramid network (FPN) [5], the multi-level features from each block of VGG16 are subsequently processed. Forcing the network to recognize all possible objects in the scenario, the object-agnostic segmentation algorithm $\mathcal{S}$ is adopted to conduct panoramic segmentation on the input features $\boldsymbol{I_l}$, where l is the layer index of the multi-scale features. SLIC, Felzenszwalb, and other unsupervised schemes are available. Then, the tensor decomposition $\mathcal{J}$ is conducted on the segmentation results to obtain object distribution features $\boldsymbol{F_D} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, \dots, \boldsymbol{f_{N_{max}}}\}$, $\boldsymbol{f_i}$ is the vector representing the existing space of i th possible object, $\mathrm{N_{max}}$ is the maximum number of segmented objects.

Emphasizing the relationship between the possible objects and image features, the matrix multiplication operation × is adopted to generate the attention matrix $\boldsymbol{F_A}$. It concentrates on the correlation between individual objects and feature channels. After calculation and convolution, the updated features $\boldsymbol{O}$ are generated, as formulated below,

$$\boldsymbol{F_D} = \mathcal{S}(\mathcal{H}(\boldsymbol{I_l}; \theta); N_{max}) \tag{1}$$

$$\boldsymbol{F_A} = \boldsymbol{F_D} \times \mathcal{H}(\boldsymbol{I_l}; \theta) \tag{2}$$

$$\boldsymbol{O} = \mathcal{H}\{(\boldsymbol{F_A} \times \boldsymbol{F_D}) \otimes \mathcal{H}(\boldsymbol{I_l}; \theta); \theta\} \tag{3}$$

where $\mathcal{H}$ represents the convolution operation, θ and ζ are the learnable parameters, ⊗ represents concatenation operation between tensors, and $\boldsymbol{O}$ is the updated features.

Preparing for mapping the object features with the known sparse point cloud features, the center regression is carried out on each $\boldsymbol{f_i}$. We take $\boldsymbol{f_i}$ as a normalized grayscale matrix and use Moments ($\mathcal{M}$) to describe the positions of center $\boldsymbol{C} = \{\boldsymbol{c_1}, \dots, \boldsymbol{c_{N_{max}}}\} = \{(x_1, y_1), \dots, (x_{N_{max}}, y_{N_{max}})\}$. Then the object eigenvector set $\boldsymbol{V} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_{N_{max}}}\}$ is acquired according to the spatial-interpolation features of $\boldsymbol{O}$ and $\boldsymbol{C}$.

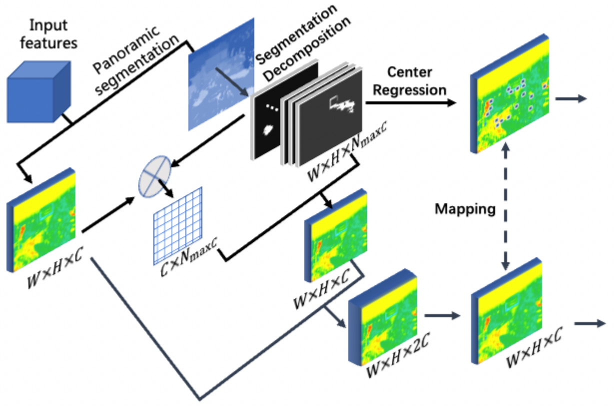

Fig. 3 Diagram of the feature extraction and enhancement module

2.2. Multi-feature Correlation Module

This module takes the known sparse point cloud $\boldsymbol{P}$, the updated features $\boldsymbol{O}$, and the corresponding center information $\boldsymbol{C}$ and $\boldsymbol{V}$ as inputs. The correlation analysis of multi-features is carried out at two levels, one is within the area of each object, the other is among all objects in the whole scene.

Firstly, the 3D sparse point cloud is projected $\mathcal{P}$ from the world coordinate to the relative image coordinate using the camera parameters. Points in the point cloud belong to the same possible object region ($\boldsymbol{f_i}$) are refined into the set $\boldsymbol{P_i} = \{\boldsymbol{p_i^1}, \dots, \boldsymbol{p_i^n}\} = \{(p_1^x, p_1^y), \dots, (p_n^x, p_n^y)\}$, which corresponding feature tensor is $\boldsymbol{F_{P_i}} = \{\boldsymbol{f_{P_i}^1}, \dots, \boldsymbol{f_{P_i}^n}\}$, where (p^x, p^y) is the position of the point on feature map after linear interpolation, n is the number of points in the i th cluster. The correlation between points within each set is processed through GCN operation $\mathcal{G}$. Center offset vectors are obtained through global average pooling operation Θ_{GAP} , which is then used for updating the features of the center $\boldsymbol{f_{P_i}^c}$. The updated center and point cloud features are fused again with the elementwise multiplication $\odot$ to update the local point cloud features. The overall object-level analysis is formulated as below,

$$\boldsymbol{f}_{\boldsymbol{P}_i}^{\boldsymbol{c}} = \Theta_{GAP}\left(\mathcal{G}(\boldsymbol{F}_{\boldsymbol{P}_i};\zeta)\right) \odot \boldsymbol{f}_{\boldsymbol{P}_i}^{\boldsymbol{c}} \tag{4}$$

$$\boldsymbol{F}_{\boldsymbol{P}_i} = \mathcal{G}\left(\boldsymbol{F}_{\boldsymbol{P}_i} \odot \boldsymbol{f}_{\boldsymbol{P}_i}^{\boldsymbol{c}};\zeta\right) \tag{5}$$

Considering the uncertainty of segmentation results, the updated center vectors $\boldsymbol{f}_{\boldsymbol{P}_i}^{\boldsymbol{c}}$ are also projected for scene-level analysis. The information such as relative distance between object centers and their similarity is applied for merging redundant objects and strengthening the difference between objects. Therein, the similarity of objects is measured by vector cosine inner product. Then, the output results are re-mapped into $\boldsymbol{O}$ for feature optimization. Finally, based on the above feature results, the bundle adjustment method is used to carry out the iterative regression of the camera pose.

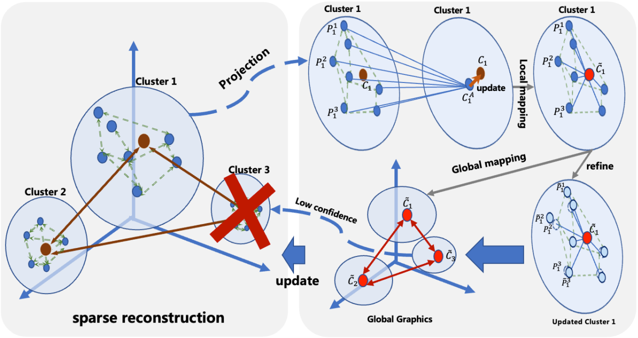

Fig. 4 Diagram of the multi-feature correlation module

3. TRAINING AND PREDICTION

During training and prediction, the bundle adjustment is adopted to solve camera pose $\langle R_l, t_l\rangle$. Our approach is trained by comparing the poses estimated at different feature levels. The goal of bundle adjustment is to minimize reprojection error. Since the nonlinear rotation matrix is introduced in pose estimation, we use the Jacobian matrix for linearization. Besides, the Levenberg-Marquardt (LM) algorithm is adopted to transform the initial estimation poses to the optimized pose through minimizing the pose error function through iteration. The optimization process follows [6].

In terms of the training, two loss functions are designed. $\mathcal{L}_{2d-f}$ focuses on features between 2D images; $\mathcal{L}_{3d-d}$ considers the mapping accuracy of 3D points under real and predicted poses, as formulated in Equation (6).

In $\mathcal{L}_{2d-f}$, it is assumed that the pose $\langle R_l^{\mathbb{0}}, t_l^{\mathbb{0}}\rangle$ of the first image is the original world pose. The features $\boldsymbol{F}^{\boldsymbol{I}_2}{}_{\mathcal{P}(R_l\boldsymbol{p}_i+t_l)}$ of each point in sparse point cloud in the second image can be obtained through 3D-2D projection with the predicted pose $\langle R_l, t_l\rangle$. Similarly, we can get the corresponding point cloud features of the first images, $\boldsymbol{F}^{\boldsymbol{I}_1}{}_{\mathcal{P}(R_l^{\mathbb{0}}\boldsymbol{p}_i+t_l^{\mathbb{0}})}$. The distance between features is defined by the L2 norm loss function, which is applied to minimize feature differences of the same point in different images and reduce the domain differences.

Besides, in $\mathcal{L}_{3d-d}$, the pose loss of the predicted $\langle R_l, t_l\rangle$ and the ground truth $\langle \widehat{R_l}, \widehat{t_l}\rangle$ are considered. After projecting the 3D spatial point into 2D space, the network parameters are optimized by minimizing the distances between projected points. Huber loss (h) is less sensitive to outliers, so it is used in Equation (8). Specific losses are as follows:

$$\mathcal{L} = \mathcal{L}_{2d-f} + \mathcal{L}_{3d-d} \tag{6}$$

$$\mathcal{L}_{2d-f} = \frac{1}{L}\sum_{l}\sum_{i}\left[\!\left[\boldsymbol{F}^{\boldsymbol{I}_1}{}_{\mathcal{P}(R_l^{\mathbb{0}}\boldsymbol{p}_i+t_l^{\mathbb{0}})} - \boldsymbol{F}^{\boldsymbol{I}_2}{}_{\mathcal{P}(R_l\boldsymbol{p}_i+t_l)}\right]\!\right]_2 \tag{7}$$

$$\mathcal{L}_{3d-d} = \frac{1}{L}\sum_{l}\sum_{i}\left[\!\left[\mathcal{P}(R_l\boldsymbol{p}_i + t_l) - \mathcal{P}\left(\widehat{R}_l\boldsymbol{p}_i + \widehat{t_l}\right)\right]\!\right]_h \tag{8}$$

During prediction, not only the camera pose, but also the centers with high confidence are predicted. The center points are projected into the 3D spatial through the predicted poses, and the 3D points will be supplemented into the sparse point cloud reconstruction data. It is believed that the small object missing problem (as shown in Fig.1) in the sparse point clouds will be ameliorated if we replenish some credible key points into the raw point cloud set. It is convinced that good sparse reconstruction results provide a better data basis for subsequent dense reconstruction.

4. EXPERIMENTS

In order to verify the generalization of the network, the Extended CMU Seasons dataset [7] and the Simulation UAV dataset are used for evaluation.

Extended CMU Seasons data was acquired by the vehicle-mounted camera, which includes scenes about Urban, Suburban, and Park. The dataset provides reference 3D reconstruction by SFM, and 50% of the data provides reference poses. In addition, about the UAV dataset, we used 3D simulation software to simulate UAV image sequences about the wide-area outdoor scene. We calibrated the internal and external parameters of the simulated camera through the chessboard calibration algorithm, and then estimated the visual depth and sparse point cloud through the COLMAP pipeline. The examples of the dataset are shown in (a) of Fig.5.

Table I Performance comparison of image retrieval

Dataset	**Method**	**Data Scene**		
		Urban	Suburban	Park
Extended CMU Seasons	D2-Net[3]	89.3	75.1	51.4
	PixLoc[4]	90.4	81.1	62.5
	Ours	**91.1**	**83.4**	**65.6**
Simulation UAV dataset	D2-Net[3]		90.5	
	PixLoc[4]		93.8	
	Ours		**94.3**	

In the experiment, in accord with pixLoc [4], the inputs are resized to 512*512, and the learning rate is $1e^{-5}$. We compared the performance of D2-Net, pixLoc, and our network. That is, we analyzed the accuracy of pose estimation by comparing the image retrieval (IR) under the location deviation threshold ⟨0.5 m, 5°⟩, as shown in Table I.

In the Suburban and Park scenes of Extended CMU Seasons, the performance of our method is superior but D2-Net's is deficient. The situation is that the camera is of low viewing angle, and the similar objects are closely arranged and occupy large image space (like trees). The feature matching method D2-Net may be forceless to it. In contrast to matching the indistinguishable features, the geometric calculation becomes the bottleneck factor for pose estimation. Besides, even though there is little difference between the three methods in the urban scene, our network and pixLoc get better results, which reveals our networks are cognitively capable of diversiform scenarios.

For the Simulation UAV Dataset, the ground truth annotations are acquired by estimating and sampling. Our network performs best in the dataset. It is worth noting that D2-Net does not perform well. In fact, because of the perspective brought by the aerial shot, the change ratios of optical flow in the foreground and far-end background are inconsistent within the UAV images. Based on the experience of existing registration algorithms, a large number of feature matching points are located in the far-end background where tiny changes happened, and insufficient attention is paid to the real foreground objects. Our method, which uses object-agnostic segmentation, reduces the pressure of registration in the background, therefore it works better.

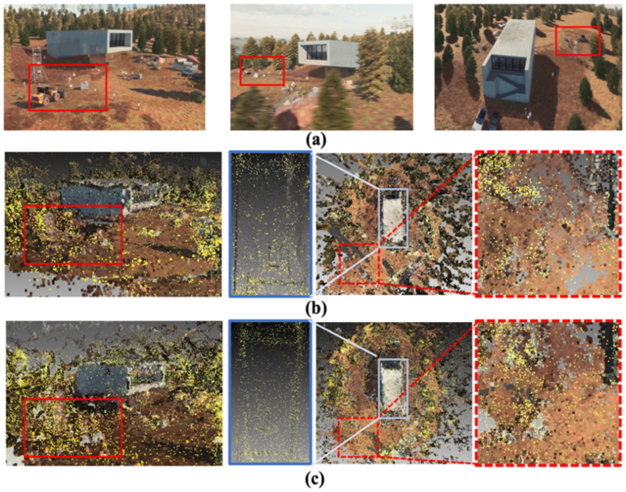

Fig. 5 (a) Examples of color images. (b) Sparse reconstruction results based on SFM and dense reconstruction results based on patch-based MVS. (c) Sparse reconstruction results supplemented with our network and dense reconstruction based on patch-based MVS.

In addition, for proving our network paying more attention to small objects, we use 3D point cloud visualization to observe the point set which is re-mapped from the 3D point predicted by our network, as shown in (c) of Fig. 5. In the red box, the neglected vehicles receive more attention; in the blue box, the distribution of the reconstructed point cloud is more uniform. This proves that our method ameliorates the situation of big object deviation and small object missing.

5. Conclusion

In this paper, a pose estimation network based on CNN and 3D geometry theory is proposed for the camera pose estimation task for UAV images of the outdoor scene. The sparse point cloud reconstructed by SFM is used as the auxiliary information to obtain the camera pose of 2D images. Notably, this network uses an object-agnostic segmentation module to extract all possible object areas in the scene, improving the attention to the features of small objects. By comprehensively processing the sparse point cloud information with the object center information through the GCN, the feature extraction process is optimized. Finally, the bundle adjustment is adopted to get the camera pose. In the existing benchmark dataset CMU and our Simulation UAV dataset, this network shows a similar effect to other pose estimation networks, especially performing better in aerial shooting UAV images of the outdoor scene.

6. ACKNOWLEDGEMENTS

This work is funded by the Open Topic of Hunan Key Laboratory of Remote Sensing Monitoring of Ecological Environment in Dongting Lake area (No:DTH Key Lab.2022-005) and supported by BUPT Excellent Ph.D. Students Foundation.

7. REFERENCES

[1] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *CVPR 2011*, Colorado Springs, CO, USA, Jun. 2011, pp. 2969–2976.

[2] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 920–929.

[3] M. Dusmanu *et al.*, "D2-Net: A Trainable CNN for Joint Description and Detection of Local Features," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8084–8093.

[4] P.-E. Sarlin *et al.*, "Back to the Feature: Learning Robust Camera Localization from Pixels to Pose," *arXiv:2103.09213 [cs]*, Apr. 2021, Accessed: Jan. 12, 2022.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," *arXiv:1612.03144 [cs]*, Apr. 2017, Accessed: Feb. 02, 2021.

[6] D. W. Marquardt, "An algorithm for least-squares estimation of non-linear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[7] T. Sattler *et al.*, "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions," *arXiv:1707.09092 [cs]*, Apr. 2018, Accessed: Jan. 12, 2022.