BeBaOI: Benchmark and Baseline Methods for Determining Overlapping Images

ISPRS Scientific Initiatives 2023

Final Report

Xin Wang (PI), Wuhan University

Yu Feng (PI), Technical University of Munich

Ronny Hänsch (Co-PI), German Aerospace Center (DLR)

Zongqian Zhan (Co-PI), Wuhan University

Minglei Li (Co-PI), Nanjing University of Aeronautics and Astronautics

Michael Gruber (Co-PI), Vexcel Imaging GmbH

Christian Heipke (Co-PI), Leibniz University Hannover

Abstract

Efficient determination of overlapping image pairs is very crucial for large scale SfM (Structure from Motion) or image orientation. This project, BeBaDOI, addresses this challenge via exploring the possibility of learning-based methods. First, a benchmark (BeDOI) with photogrammetric referenced overlapping relationship is published, which includes more than 13k images of various scenarios. Second, based on several popular backbones and our BeDOI, a supervised fine-tuning solution is presented with triplet loss for the generalization on overlapping image pair determination. In addition, to further speed up the retrieval, an unsupervised method is introduced to build an efficient hierarchical vocabulary three. Finally, this complete solution is successfully applied on both large-scale offline SfM and online SfM. In this context, two conference papers (GSW 2023 and ECCV workshop 2024) are published, one journal paper is already accepted by the PFG-Journal and one journal paper was submitted to the ISPRS Journal P&RS, all the pre-trained backbone and benchmark are online available. In the future, we hope this project can further benefit the community in large-scale SfM, VSLAM, 3D reconstruction.

Motivations, Objectives and Partnerships

SfM (Structure from Motion) addresses the problem of estimating the image poses and the corresponding sparse object 3D points, which is in general identical to the basic goal of image orientation (Wang et al., 2019). Over the past decade, SfM has obtained ample achievements, especially for large-scale image datasets (Frahm et al., 2010; Wilson and Snavely, 2014; Zhu et al., 2018; Schonberger and Frahm, 2016), thanks to some popular open packages (such as, Colmap (Schonberger and Frahm, 2016), OpenMVG (Moulon et al., 2016) etc.). However, a challenging problem of matching visual overlapping image pairs is posed when dealing with very large image datasets, such as crowdsourced images of various landmarks and images collected from social media (Flickr, Instagram, etc.), or even images taken by professional photogrammetrists. For the crowdsourced datasets, images were taken by various tourists with unidentical cameras in an arbitrary manner, which makes the images sorted in an unordered way. In the field of photogrammetry, with the knowledge of GPS and IMU, images that are spatially closed can be easily determined and image matching is only carried out on these closed pairs. However, GPS and IMU cannot be applied everywhere and their corresponding signal is often obstructed by trees and buildings in urban areas (Goforth and Lucey, 2019), thus, the common close-range images without GPS/IMU taken in an arbitrary way are also unordered.

One intuitive idea to handle unordered images is exhaustive image matching, i.e., matching every possible pair. However, it becomes impractical for just several hundreds of images as the complexity grows quadratically with the number of images, i.e., $N(N\text{-}1)/2$ image matchings for N images (Wang et al., 2019). To accelerate the image matching procedure, various methods for designing an efficient indexing structure were proposed, recently, most of the SfM systems adopt the Bag-of-Word methods (Sivic and Zisserman, 2003; Nister and Stewemius, 2006) which aggregate local features and describe an image via an aggregated global vector. Overlapping image pairs are found by comparing the global features. Similarly, Visual vocabulary tree is built to search for the nearest neighbors of local features (Havlena and Schindler, 2014). Both BoW and VoC take handcrafted local features (SIFT (Lowe, 2004), ORB (Rublee et al., 2011)) as input, which can be directly fed into subsequent geometric processing. This might be one of the reasons why they are generally favored in SfM pipelines. Despite the popularity of BoW and VoC, due to some pre-setting free parameters (e.g., the number of bags and clusters, or the depth of the VoC), their retrieval efficiency and accuracy are limited and decrease as the number of images increases.

To guarantee the scalability of retrieval, CNN-based methods that have shown superior performance on object image retrieval (Philbin et al., 2007; Philbin et al., 2008; Chen et al., 2021) (i.e., to distinguish whether the target image's content is similar to a cat or a dog) come into the notice of determining visual overlapping pairs, Tolias et al. (2016) and Radenovic et al. (2016) employed the feature maps of several renowned pre-trained CNN architectures to yield a compact global feature, and similar image pairs are identified by investigating the distances of two images

in the global latent feature space. In this case, the corresponding time efficiency is improved since each image is represented by just one global feature vector. In the context of SfM, visual overlapping image pair implies that two images observe the same area in 3D object space, but some parts of the region may not be identical due to the changes in viewing positions. This in principle differs from object image retrieval that identifies semantically similar images, in which the corresponding CNN-based methods (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015) were trained using relevant training datasets (e.g., ImageNet (Deng et al., 2009)). In addition, exhaustive comparison between global features is yet not the most efficient solution. These may result in the CNN-based method not being successfully used in mainstream SfM and SLAM solutions (Mur-Artal et al., 2015; Engel et al., 2014). Motivated by the high time efficiency and scalability of learning-based methods, it should be advocated to further explore the possibility of CNN-based methods in the application of SfM. In this project, three objectives have been achieved:

- To cope with scarce benchmarks and the mentioned domain gap, we provide a benchmark with geometrically correct references of overlapping image relationships - BeDOI, including 13,667 images of several different content (such as urban buildings, countryside, forest, etc). It cannot only be applied for evaluating performance of relevant overlapping image pairs retrieval algorithms, but also cast as training data for learning-based global feature extractors to boost the sensitivity for pairwise overlapping information;

- Based on the generated BeDOI, we present a simple yet efficient fine-tuning solution that are supposed to extend learning-based backbones' generality for detecting overlapping image pairs with various rotations;

- An efficient indexing solution of hierarchical vocabulary tree is proposed with fine-tuned global features, which can further improve the time efficiency of image retrieval.

This project has been completed thanks to the collaboration of researchers from five well-known academic institutions and one world-renowned industry company: School of Geodesy and Geomatics (Wuhan University, China), Chair of Cartography and Visual Analytics (Technical University of Munich, Germany), Microwaves and Radar Institute (German Aerospace Center (DLR), Germany), Institute of Photogrammetry and GeoInformation (Leibniz University Hannover, Germany), College of Electronic and Information Engineering (Nanjing University of Aeronautics and Astronautics, China) and Vexcel Imaging GmbH (Austria).

Benchmark data collection and generation

In this section, we first give an overview introduction of the benchmark - BeDOI dataset. Then, the automatic procedure for generating BeDOI is explained.

Introduction of BeDOI

In general, BeDOI is composed of 11 high-resolution image datasets, including UAV images captured via a nadir camera and oblique photogrammetric images with multiple cameras, as well as manually self-collected close-range images with different overlap degrees, which is tailored for overlapping image pair identification on photogrammetric image datasets. More specifically, as Tab.1 lists, in total, 13,667 images covering various categories of areas are collected, such as urban buildings, woodland, countryside, scenic spots, etc. Fig. 1 shows several examples.

Name	Image Num.	Source	Category
SKFX	60	Close range	Historic Relics
GB	68	UAV	Scenic Spot
GRAZ	250	Oblique	Urban City
YD	374	UAV	Scenic Spot
NH	606	UAV	Building
TZH	1060	UAV	Countryside
SXKQ	1185	UAV	Forest
JYYL	1429	Close range	Building
XHSD	2133	Oblique	Urban City
WHU	2652	UAV	University
SHHY	3850	Oblique	Village
BeDOI	13667	Multi-sources	Multi-categories

Table 1. Information of each dataset in BeDOI.

(a) Scenic spot (b) Urban area

(c) Cultural relic (d) Forest

Figure 1. Example images of BeDOI.

Automatic annotation for generating BeDOI

The overall pipeline to automatically generate BeDOI is illustrated in Fig. 2, in which pre-processing is for obtaining 3D mesh model and image orientation parameters, and automatic annotation is for estimating referenced overlapping relationships:

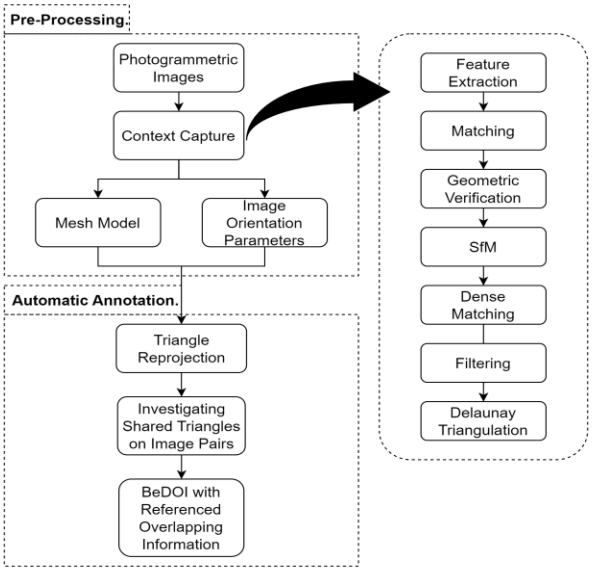


Figure 2. Flowchart of BeDOI generation.

Pre-Processing. Given a set of collected images, this step is to generate corresponding photogrammetric information, i.e., 3D mesh models and image orientation parameters. Following the canonical photogrammetric processing, several consecutive procedures are required: feature extraction and matching, SfM, stereo dense matching and multi-view fusion, filtering and 3D mesh construction (including Delaunay triangulation, texture re-organization etc.). Note that orientation information is computed after SfM. This BeDOI processing chain might seem counterintuitive since image matching is usually completed before the 3D mesh model is built. However, leveraging a 3D mesh model for identifying real overlapping image pairs is not only a viable but also highly advantageous solution, as most local features are typically not invariant to large view angle change, e.g., oblique images. Such a procedure is beneficial even for state-of-the-art learning-based local feature extractors can only slight improve the matching performance (Yi et al., 2016). This motivates us to explore 3D mesh models for estimating correct overlapping information in a geometrically rigorous manner. One sample mesh model of JYYL is shown in Fig. 3.

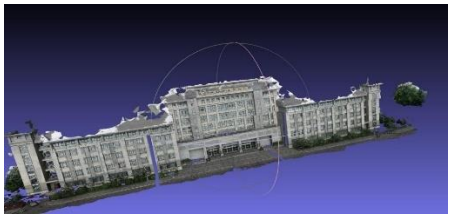


Figure 3. 3D mesh model of JYYL.

Automatic Annotation. Based on the collinearity equation, we present an automatic annotation method for geometrically correct referenced overlapping image pairs using the generated 3D mesh model and image orientation parameters. The basic idea is to reproject every triangle on every image. Shared triangles between two images are explored for determining the corresponding overlapping degree. The more common reprojected triangles, the larger the corresponding overlapping area will be. To estimate accurate triangle reprojection, it is necessary to deal with occlusions. Fig. 4 shows that there are many incorrectly identified overlapping areas without occlusion detection which can lead to incorrect results in BeDOI. In this work, occlusion is detected by the number of triangles that the corresponding ray (from the camera center to the center of the target triangle) passes through. No occlusion happens if and only if the number is zero. Furthermore, in order to enhance occlusion detection speed, we construct an AABB tree for the mesh model. After the occlusion detection, the correct triangle information of the image can be obtained.

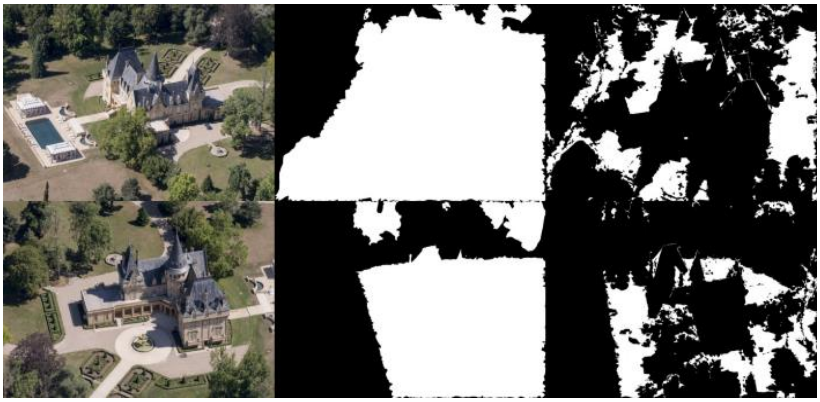


Image pair WoO WO

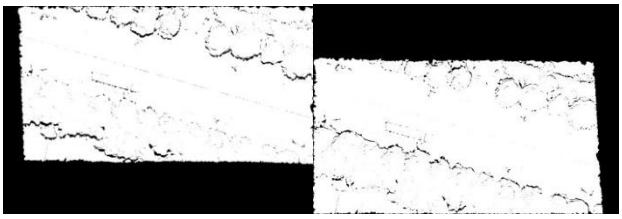
Figure 4. With occlusion (WO) vs. Without occlusion (WoO). White pixels indicate the overlapping area via the proposed triangle reprojections.

After triangle reprojection and occlusion detection, the similar or overlapping degree of image pair (i, j) can be computed as follows:

$$OI_{ij} = \sqrt{\frac{|TR(i) \cap TR(j)|_n}{|TR(i)|_n} \cdot \frac{|TR(i) \cap TR(j)|_n}{|TR(j)|_n}} \tag{1}$$

where $|.|_n$ returns the number of triangles, TR(i) ∩ TR(j) represents the set of triangles that can be observed in both image i and j. Straightforwardly, the larger the value of OI_{ij} is, the more similar the image pair (i, j) is. Based on the conventional photogrammetric regularity, image pairs i and j can be identified as overlapping if OI_{ij} values exceed 0.3. Fig. 5 qualitatively shows the determined overlapping region, where the highlighted part in Fig. 5(a) is the overlapping area of the two images, Fig. 5(b) is a binary image with the white region corresponding to the highlighted area Fig. 5(a).

(a) Determined overlapping region. Highlighted parts are overlapping area



(b) Binary results of overlapping region. White regions indicate overlapping area.

Figure 5. Qualitative results of determined overlapping region.

Ultimately, the overlap or similarity degree among all image pairs can be calculated by equation (1). In this paper, we sorted the values of OI_{ij} in descending order based on the number of overlapping patches. For a binary classification, image pairs with OI_{ij} values exceeding 0.3 are the referenced overlapping ones.

Learning-based baseline method for determining overlapping images

In this section, we provide more details of the proposed method for identifying overlapping image pairs and the corresponding offline and online retrieval mode. Four parts are included: 1. General overview of the proposed work; 2. Supervised backbone fine-tuning solution; 3. Unsupervised indexing structure of hierarchical vocabulary tree; 4) Offline and Online OIP retrieval.

Overview of the proposed baseline method for detecting overlapping images

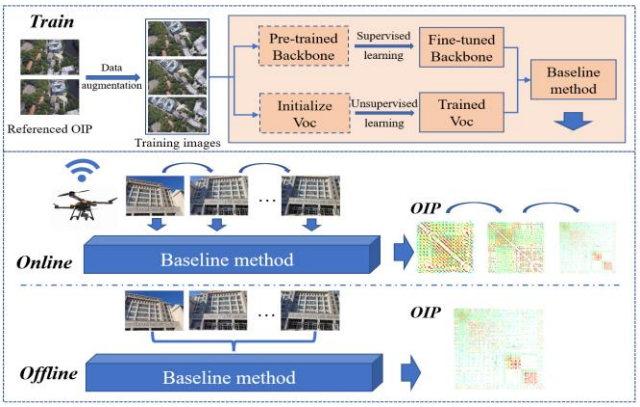


Figure 6. Overview of the proposed baseline method for overlapping images.

Fig. 6 illustrates the overall framework of our method, comprising two phases: training and application.

Training Phase:

1. We begin with pre-trained backbones and perform supervised fine-tuning using referenced overlapping image pairs to ensure the extracted global features are sensitive to overlapping information.

2. An unsupervised process generates a lookup dictionary of a hierarchical vocabulary tree (Voc) tailored to the fine-tuned global features.

Application Phase:

1. Offline Retrieval. All images are processed together. Their global features are extracted, and the whole overlapping relationships are determined using the hierarchical vocabulary tree.

2. Online Retrieval. This is analogous to dynamic searching procedure. Each new image's global feature is extracted, which is then used to traverse on the constructed hierarchical vocabulary tree for finding overlapping images in the database.

Supervised backbone fine-tuning

Introduction and augmentation of training dataset. In this paper, we employ the benchmark of BeDOI (Zhan et al. 2023) to generate training samples for our fine-tuning solution, the reasons are: first, **diverse sources and scenarios.** BeDOI contains photogrammetric images from various sources (close range, UAV, oblique) and scenarios (urban, countryside, village, forest), enhancing the generalization of the fine-tuned backbones in handling diverse photogrammetric images. Second, referenced overlapping relationships. BeDOI provides referenced overlapping relationships between images, which is crucial for our fine-tuning purpose. We utilize full photogrammetric information to rigorously determine overlapping image pairs. Specifically, commercial software is used to generate 3D mesh models and image orientations. Mesh triangles are reprojected onto the image space with occlusion detection, and overlapping image pairs are identified based on the number and area of reprojected common mesh triangles.

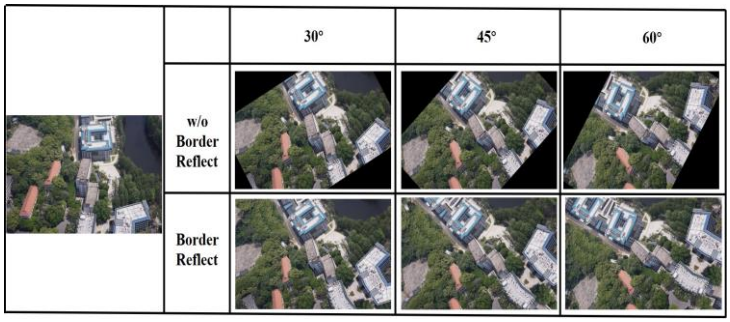


Figure 7. Performance of BORDER_REFLECT.

To further improve the model's generalization ability, we employ data augmentation to simulate diversity and noise in real-world scenarios. This strategy involves random rigid transformations including rotation, flipping, scaling, and brightness changes, which benefit the backbone model

become sensitive to different perspectives and transformations. In particular, to avoid the degradation of the black edges and pixel loss caused by simulated rotation, the BORDER_REFLECT (BR) technique. This method applies mirror reflection to replicate the pixel information on the sides of the region of interest (RoI), effectively filling in the lost pixels as shown in Fig. 7. The open-sourced OpenCV package is utilized for implementing BR in this project.

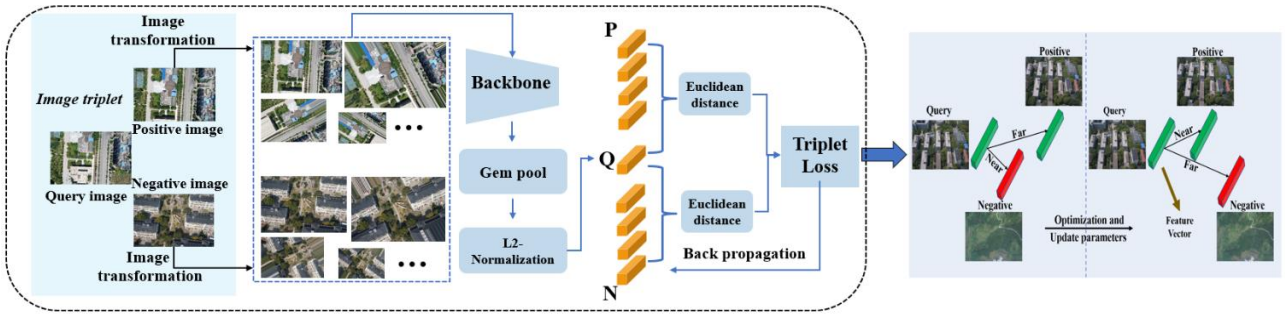


Figure 8. Backbone fine tuning using triplet images and triplet loss.

Backbone Fine-tuning. The inclusion of hard negative samples (non-matching local features, dissimilar image pairs) enhances the performance of feature matching and object retrieval. In our case, these hard negatives are non-overlapping image pairs. As illustrated in Fig. 8, we input triplets comprising positive overlapping image pairs and negative non-overlapping image pairs into a shared-weight triplet backbone model. The primary goal is to refine the pre-trained weights to distinguish between overlapping and non-overlapping image pairs. Our fine-tuning solution involves the following components:

1. Backbone encode layer. For the feasibility of triple images, a three-branch backbone encode with same weights is fine tuned. In this study, five popular backbones (VGG16, ResNet101, GoogleNet, SwinT and PvT) are used as examples to demonstrate the efficacy of the proposed fine-tuning solution. In addition, as Jun et al. (2019) suggests, FC (fully connected) layer typically yields inferior image retrieval results due to a lack of geometric invariance and spatial information. Therefore, we retain only the convolution layers and transformer blocks as encoders. Then, based on the cropped backbones, we can easily obtain activations of multi-channel feature maps. In general, these activations are with very high dimension, they are unsuitable and time inefficient for retrieval tasks. Thus, an aggregation layer is required to generate concise global features. L2 normalization is applied to channel-wise activations before the aggregation layer.

2. Aggregation layer. To generate compact and effective global feature, in this work, the widely-used GeM pooling (Radenović et al. 2019) is embedded as aggregation layer, as shown in Fig. 9 and equation (2). Here $x_{i,j,c}$ is feature value from the location (i, j) of c-th feature map, H and W are the height and width of feature map, p denotes the power index and can be refined during training. The power operation with parameter p is perform on the elements of each feature, then channel wise averaging pooling is followed for c-th dimensional feature. Finally, another 1/p-th power operation is applied to generate GeM feature. After *L2* normalization, we obtain the

standardized vector as the final feature descriptor.

$$y_c = \left(\frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(x_{i,j,c}\right)^p\right)^{\frac{1}{p}} \quad (2)$$

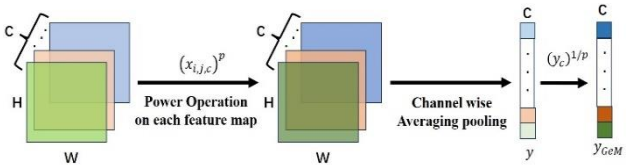


Figure 9. GeM Pooling workflow.

3. Training with triplet loss. While Siamese networks (Radenovic et al. 2016) and pairwise loss (such as contractive loss) have been successfully used to refine backbones, triplet loss is yet considered in our fine-tuning solution as it has better performance on avoiding overfitting (Hou et al., 2023). To measure similarity between two images (I_i, I_j), we first extract the global feature *GF* using the mentioned backbone encode and aggregation layers, and compute Euclidean distance $D(I_i, I_j)$ to estimate similarity of two images. Before training, tuples $\{T(I_Q, I_P, I_N)\}$ that contain positive image pair (I_Q, I_P) and negative image pair (I_Q, I_N) are selected. Positive pairs are selected among all referenced overlapping image pairs in BeDOI, and the third non-overlapping image I_N is randomly selected from other sub-datasets. In addition, to extend model generalization, the proposed augmentation methods are randomly performed on I_P and I_N. Finally, with a pre-set constant margin *M*, the conventional triple loss (Schroff et al. 2015) given by (3) is employed in our fine tuning. The goal is to fine-tune the backbone so that the distance between I_Q and I_P is minimized, while the distance between I_Q and I_N is maximized to at least margin *M*.

$$Loss(T(I_Q, I_P, I_N)) = max\left(D\left(GF_{I_Q}, GF_{I_P}\right) + M - D\left(GF_{I_Q}, GF_{I_N}\right), 0\right) \quad (3)$$

Unsupervised indexing structure of hierarchical vocabulary tree

While global features with exhaustive pairwise comparison can speed up image matching (Hou et al. 2023), we propose a lookup structure based on a hierarchical vocabulary tree tailored for global features, which can further improve the time efficiency of retrieval and the image matching. Our hierarchal vocabulary tree is trained based on the fine-tuned global features of images from BeDOI.

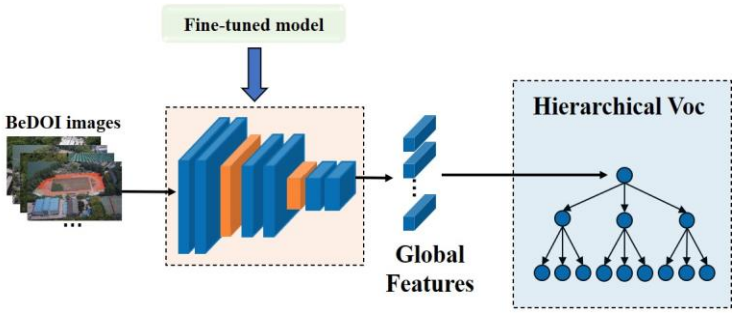


Figure 10. Construction of Hierarchical vocabulary tree.

Construction of hierarchical vocabulary tree. In this work, the construction of hierarchical vocabulary tree is very simple and illustrated by Fig. 10. First, all the training images of BeDOI are used, and their global features are extracted using fine-tuned backbones. These global features are then clustered into k sub-nodes using a canonical unsupervised classification method – K-means[1], where k is the pre-set number of sub-clusters for each node. The features within each sub-node are recursively clustered into further sub-nodes using K-means until a pre-set layer L is achieved, or the sub-node contains fewer than k global features. The resulting sub-nodes are recorded to facilitate fast retrieval.

Discussion. The key motivation behind the hierarchical vocabulary tree is that overlapping images with similar global features should be grouped into the same sub-nodes, aiding in efficient feature search. However, the depth (L) and width (k) of the vocabulary tree can affect both retrieval accuracy and efficiency. If the size of the hierarchical vocabulary tree is too large, it typically has superior retrieval precision as the feature space are split more elaborately, but the storage memory and searching time increase, because more sub-nodes are needed to be stored and traversed. Conversely, for a small hierarchical vocabulary tree, the retrieval speed and memory can be efficient, but it may contribute to a coarse split feature space which could return ambiguous candidate similar images and affect retrieval precision. On the other hand, for datasets of varying sizes, a relevant applicable hierarchical vocabulary tree should be advocated, the ideal complexity for hierarchical vocabulary tree to search one query is $O(L \times k)$, which should be smaller than $O(n)$, n is the number of images[2]. To balance retrieval speed and precision, the vocabulary tree should be designed to ensure time-efficient retrieval while maintaining good accuracy, which implies that the structure of the vocabulary tree should be with careful consideration.

Offline & Online OIP retrieval

Based on the fine-tuned backbone model and hierarchical vocabulary tree, a baseline method for determining overlapping image pairs can be easily expected with the improved image representation and efficient indexing solution. This section presents two retrieval applications of online and offline working mode, as Fig. 9 shows. The key assumption is that overlapping images are supposed to fall into the same nodes of hierarchical vocabulary tree, i.e., similar global features should be contained by the same sub-cluster.

Offline OIP retrieval. Offline retrieval mode considers all collected images together and returns all the potential overlapping image pairs at one time, which is just suitable for improving matching speed of conventional offline SfM (Hou et al. 2023). All extracted global features are sent into the constructed hierarchical vocabulary tree, traversing from the top to the bottom layer. The nearest sub-node for each global feature is determined by comparing the Euclidean distances between sub-cluster centers and the corresponding global feature. Images within the same sub-node are considered overlapping pairs. If a query image falls into a sub-node with only a few

[1] In this work, K-means is implemented by using the "KMeans" module from the pytorch package of sklearn.
[2] When traversing on the hierarchical vocabulary tree, each query has to be compared with $L \times k$ sub-nodes to determine which sub-nodes it should be clustered into.

candidate images (e.g., fewer than 30), its sister sub-nodes (inheriting from the same parent sub-node of the previous layer) are explored until sufficient candidate overlapping images are found.

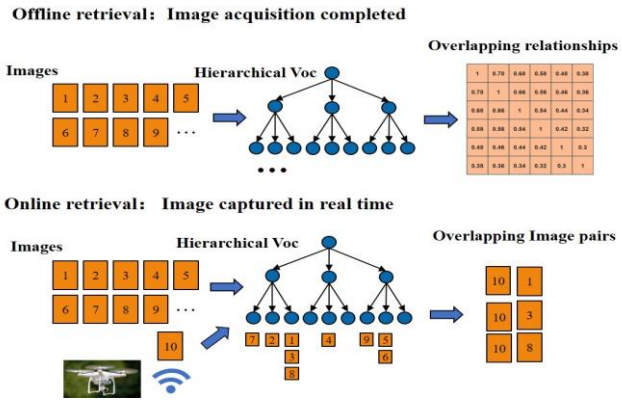


Figure 11. Offline and Online retrieval based on the proposed baseline method.

Online OIP retrieval. Unlike the offline OIP, which processes all captured images together and identify the whole overlapping information, online working mode handles each newly captured image sequentially and dy-namically in real-time. The goal is to fast find new im-age's candidate overlapping images within a database, which is the most important step for online SfM (Zhan et al. 2024). In this case, the database denotes a solved pho-togrammetric block containing many already registered images. For each new image, the extracted global feature is traversed along the hierarchical vocabulary tree, similar to offline mode, the database images that stay in the same node as new image does are considered as the resulted overlapping images.

Data and Pre-trained Model Delivery

The dedicated webpage on the open public git has been implemented: (https://github.com/WHUHaoZhan/BeDOI and https://github.com/wzwcumt/LOIP-for-SfM). Any interested researcher can learn more about the our BeDOI benchmark from the first link and our learning-based baseline method for determining overlapping image pairs from the second link. The data are freely available at the following link: https://pan.baidu.com/s/1gcS4_fk52nZIWoFtczZzow (extraction code: 1234) and the pre-trained weights of pre-trained backbones and hierarchical vocabulary trees of various sizes can be downloaded via https://pan.baidu.com/s/1PG_BgpMnSgAAO4gGQUPpPA (extraction code: jquc).

Dissemination

The Scientific Initiative has been largely advertised during the ISPRS Geospatial Week 2023 in Cairo with a dedicated presentation given by the PI Xin Wang. It is still under development and a scientific paper (Wang et al., 2024) that has been accepted by the PFG-Journal of Photogrammetry,

Remote Sensing and Geoinformation Science. This paper will give more details on the learning-based baseline method for determining overlapping image pairs. In addition, the output of this project is successfully applied in the work of online SfM (Zhan et al., 2024), resulting in two scientific papers – one is submitted to the ISPRS Journal of Photogrammetry and Remote Sensing, another one is accepted by the proceedings of European Conference on Computer Vison (ECCV) workshop (Gan et al., 2024).

References

Chen, W., Liu, Y., Wang, W.P., Bakker, E.M., Georgiou, T., Fieguth, P., Liu, L., Lew, M.S., 2021. Deep Image Retrieval: A Survey. In: arXiv preprint arXiv: 2101.11282.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.F., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 248-255.

Engel, J., Schöps, T., Cremers, D., 2014. LSD-SLAM: Large-scale direct monocular SLAM. In European conference on computer vision. Springer, Cham, pp. 834-849.

Frahm, J. M., Fitegeorgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.H., Dunn, E., Lazebnik, S., Pollefeys, M., 2010. Building Rome on a cloudless day. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 368-381.

Gan, W. T., Yu, Y. F., et al., 2024. LVG-SfM: Learning-based View-Graph generation for robust on-the-fly SfM. In: Proceedings of the European Conference on Computer Vision (ECCV) workshops.

Goforth, H., Lucey, S., 2019. GPS-Denied UAV Localization using Pre-existing Satellite Imagery. In: Proceedings of the International Conference on Robotics and Automation (ICRA), pp. 2974-2980.

Havlena, M., and Schindler, K., 2014. VocMatch: Efficient Multiview correspondence for structure from motion. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 46-60.

Hou, Q., Xia, R., Zhang, J., Feng, Y., Zhan, Z., and Wang, X., 2023. Learning visual overlapping image pairs for SfM via CNN fine-tuning with photogrammetric geometry information. Int. J. Appl. Earth Obs. Geoinformation, vol. 116, p. 103162.

Jun H., Ko B., Kim Y., Kim I., and Kim J., 2019. Combination of multiple global descriptors for image retrieval. ArXiv Prepr. ArXiv190310663.

Johnson J, Douze M, and Jégou H (2019) Billion-scale similarity search with GPUs. IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547.

Krizhevsky, A., Sutskever, I. Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the Neural Information Processing Systems (NeurIPS), pp. 1097-1105.

Lowe, D. G., 2004. Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 60, pp. 91–110.

Moulon, P., Monasse, P., Perrot, R., Marlet, R., 2016. OpenMVG: Open multiple view geometry. In: Proceedings of the International Workshop on Reproducible Research in Pattern Recognition (RRPR), pp. 60-74.

Mur-Artal, R., Montiel, J.M.M., Tardos, J.D., 2015. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5), pp. 1147-1163.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2161-2168.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2007. Object retrieval withlarge vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in quantization:Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Radenović, F., Tolias, G., Chum, O., 2016. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-20.

Radenović, F., Tolias, G., and Chum, O., 2019. Fine-tuning CNN Image Retrieval with No Human Annotation. in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1655-1668.

Rublee, E., Rabaud, V., Konolige, K., Bradski G., 2011. ORB: An efficient alternative to SIFT or SURF. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.2564-2571.

Schönberger, J.L., Frahm, J.M., 2016. Structure-from-Motion Revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104-4113.

Schroff, F., Kalenichenko, D., and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823.

Sivic, J., Zisserman, A., 2003. Video google: A text retrieval approach to object matchingin videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1470-1477.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: arXiv preprint arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, P., 2015. Going Deeper with Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9.

Tolias, G., Sicre, R., Jégou, H., 2016. Particular object retrieval with integral max-pooling of CNN activations. In: Proceedings of the International Conference on Learning Representations (ICLR).

Wang, X., Rottensteiner, F., Heipke, C., 2019. Structure from Motion for ordered and unordered image sets based on random k-d forests and global pose estimation. ISPRS Journal of Photogrammetry & Remote Sensing, 147, pp. 19-41.

Wang, X., Wang, Z. W., Xu, Y. W., Zhan, Z. Q., 2024. Learning-Based Baseline Method for Efficient Determination of Overlapping Image Pairs and Its Application On both Offline and Online SfM. PFG (2024). https://doi.org/10.1007/s41064-024-00312-z.

Wilson, K., Snavely, N., 2014. Robust global translations with 1DSFM. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, pp. 61-75.

Zhan, H., et al., 2023. BEDOI: BENCHMARKS FOR DETERMINING OVERLAPPING IMAGES WITH PHOTOGRAMMETRIC INFOR-MATION. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., vol. XLVIII-1/W2-2023, pp. 1685–1692, 2023, doi: 10.5194/isprs-archives-XLVIII-1-W2-2023-1685-2023.

Zhan, Z. Q., Xia, R., Yu, Y. F., Xu, Y.B., Wang, X., 2024. On-the-Fly SfM: What you capture is What you get. ISPRS Ann. Photogramm. Re-mote Sens. Spatial Inf. Sci., X-1-2024, 297–304, https://doi.org/10.5194/isprs-annals-X-1-2024-297-2024.

Zhu, S., Zhang, R.Z., Zhou, L., Shen, T.W., Fang, T., Tan, P., Quan, L., 2018. Very Large-Scale Global SfM by Distributed Motion Averaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4568-4577.