Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 5, Issue 12: 22-28, DOI: 10.25236/AJCIS.2022.051204

A cross-view UAV image localization method based on directional feature alignment and visual transformer

Xing Liu

School of Information Science and Engineering, Chongqing JiaoTong University, Chongqing, 400074, China
779661864@qq.com

Abstract: *Using satellite images and UAV images to locate the same geographical target can provide new ideas for UAV positioning and navigation. However, the images from the two different remote sensing platforms, UAV and satellite, have a huge difference in appearance, which is a huge challenge for this task. Existing methods are usually limited to convolutional neural networks, leading to a lack of utilization of global information of images, and these methods do not focus on the spatial information of images. To address these issues, this research propose a method that extracts posture information using SFM(Structure-from-Motion) and then uses the posture to align the spatial features of the image, and introduces a visual transformer to focus the network on acquiring the common feature space shared by the viewpoint sources. In this study, a large number of experiments have been carried out on the large datum dataset University-1652. The experimental results show that the method proposed in this paper outperforms the baseline and has the same advantages as other advanced methods.*

Keywords: *Image Retrieval, Cross-View, Geo-Localization, SFM, Transformer*

1. Introduction

UAVs are widely used in various fields today. In addition to the performance of the vehicle itself, an excellent navigation and positioning system plays a crucial role in the successful completion of the mission. The current UAV navigation system usually adopts a combination of INS and GPS navigation, but the future navigation system will definitely develop in the direction of anti-jamming, high-precision, all-weather, and highly autonomous, and in order to meet this development need, the vision-assisted navigation system will definitely become a preferred technical solution by virtue of its autonomy and high accuracy. Therefore, it is of great significance to study the cross-view image geo-localization and to realize the retrieval between different source images.

The navigation and positioning system based on cross-view image geo-localization is to match the images taken by the vehicle in real time with the satellite images stored internally, and after successful retrieval, obtain the current position based on the geographic location information carried by the satellite images and feed that position information to the main navigation system of the vehicle to realize the correction of the navigation system. In this process,retrieve the satellite view image as a positioning task through the UAV view image, and retrieve the UAV view image as a target localization task through the satellite view image, as shown in Figure 1.

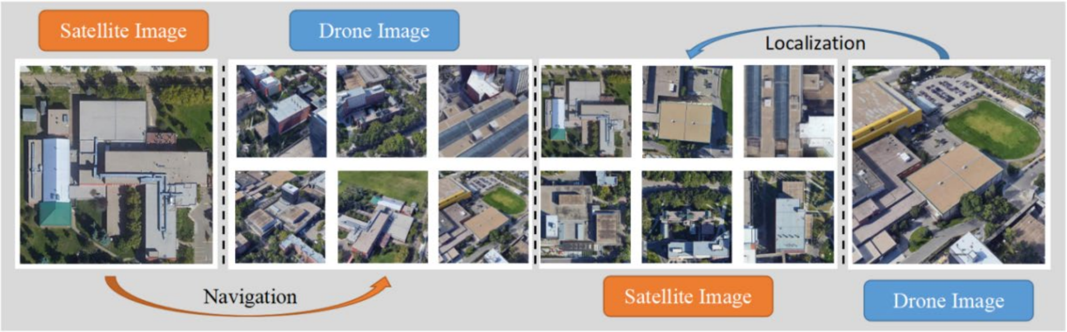

Figure 1: Cross-view geo-localization and navigation based on satellite and UAV images.

However, images from different platforms (satellites and UAVs) vary greatly in color, direction, angle, field of view and other aspects, which is the main reason for the difficulty of retrieval. What methods should be taken to bridge the gap between cross perspective images is an unsolved problem.There are three main methods for the retrieval of geographic images with great differences. The first is to use manual features, the second is to treat the cross-view geo-localization task as a classification problem, and the third method based on Metric Learning aims to learn the similarity of two images through the network. They will be described below.

1.1. Manual Features

Cross-view image retrieval is a method to retrieve images of the same scene from different views (e.g., ground, UAV, and satellite views) across views, and is the main research method for cross-view geo-localization. Its early research was mainly based on image retrieval between ground views [1]. However, the problems of severe occlusion of trees, vehicles, and pedestrians, small and limited field of view, and small coverage of ground images have led to the problem of too low efficiency in matching them with images from other views. In contrast, the satellite view with global geo-localization markers attached has irreplaceable superiority of ground view, such as no occlusion, small variation, and wide coverage. As a result, the mainstream retrieval has shifted from single ground view image matching to matching ground view with air view [2]. However, this shift brings other problems, i.e., severe spatial domain difference (domain gap). The problem of spatial domain difference of viewpoints between ground and aerial views makes it difficult for traditional manual feature methods such as SIFT (Scale-Invariant Feature Transform) [3] and SURF (Speed Up Robust Feature) [4] methods to extract complex and discriminative viewpoint invariant features, and cross-view image retrieval has a huge challenge.

1.2. Classification

The purpose of treating cross-view geo-localization task as a classification problem is to map features from different views to the same feature space for classification matching, and identification loss is generally used for training. Zheng et al. [5] regarded this as a classification task, and used three CNN branches to achieve the matching of satellite view, UAV view and ground view on their proposed dataset University-1652 [6] with category labels based on geographic objects, and adopted the example loss optimization model. The feasibility of UAV positioning and navigation task is successfully verified. Wang et al. [6] proposed the LPN method by using the adjacent areas of the satellite view image and UAV view image as auxiliary information. Specifically, using the square ring feature partition strategy not only has good scalability for the rotation changes that often occur in the drone view image, but also naturally provides different degrees of attention to different areas at the distance center. In other words, using context information without using additional estimators. Hu et al. [7] took into account the style deviation caused by camera, weather and seasonal changes, and adopted the method based on color levels to unify the style of UAV image and satellite image. In addition, they also used mesh division to carry out local feature alignment.

1.3. Metric Learning

Tian et al. [8] extracted the building by object detection method, then used the building as a bridge between ground image and aerial image, and proposed for the first time to match K most similar images through twin network for cross-view image matching, so as to achieve the goal of geographic positioning in urban environment. Hu et al. [9] use double branch convolution network respectively to extract ground image and the local characteristics of satellite image, and then use the Netvlad [10] the local characteristics of extraction for aggregation, get global description vector, then training, end-to-end similarity between two views are samples of global description vector distance minimization, maximum distance will be negative sample. Finally, according to the distance between the global description vectors of the two views, the cross-view image geo-localization of the same location is realized and the stable rank1 index performance is achieved for the first time in the cross-view geo-localization task. REGMI et al. [11] proposed an image generation method based on Conditional Generative Adversarial Nets (CGANs)[12] to reduce the visual difference between two views. In addition, they use Weighted Soft Margin Triplet Loss (WSM)[9] to assist training. Their method can generate plausible aerial views from the corresponding ground view images and then match them.

In order to solve the existing problems, such as the classification task and measurement task are separated, the three-dimensional information of geographic target is not fully used, and the sampling is not balanced. This research propose to obtain the pose information of UAV image by SFM algorithm, and use the pose information to align the orientation features of satellite image and UAV image, so as to bridge the gap between cross-view images. In addition, in order to solve the problem of unbalanced sample size, the common twin network structure is abandoned, and the common features of cross-view images are learned through the visual transformer network, and the global features of images are extracted to make full use of the context information around the target buildings.

The work of this research is mainly in the following two aspects:

- In order to help neural networks acquire geographic targets and their context information, this research provide a directional alignment strategy to transform the original image to perform the same preprocessing on the image. Based on the SFM algorithm, this study makes full use of the 3D information of UAV view images, and uses the pose information to bridge the differences in the directional features of cross-view images by means of rotating clipping, so as to enhance the unity of features.

- An improved transformer model is proposed to solve the positioning and navigation problems of UAVs. It can aggregate global coarse-grained information and local context information of satellite and UAVs images from multiple perspectives by sliding window in a unified network architecture, and learn fine-grained point-of-view invariant features. Thus, the target location image can be characterized more completely.

2. Proposed Method

2.1. Cross-View Image Orientation Alignment Strategy

Cross-view image retrieval needs to find a robust way to represent the common features of the same geographical target image, that is, to minimize the feature invariance between and within domains. The direction of the satellite image is certain, but due to the maneuverability and flexibility of the UAV, the direction of the UAV image is uncertain, which is a great difficulty to find the common feature space of the two, as shown in Figure 2.

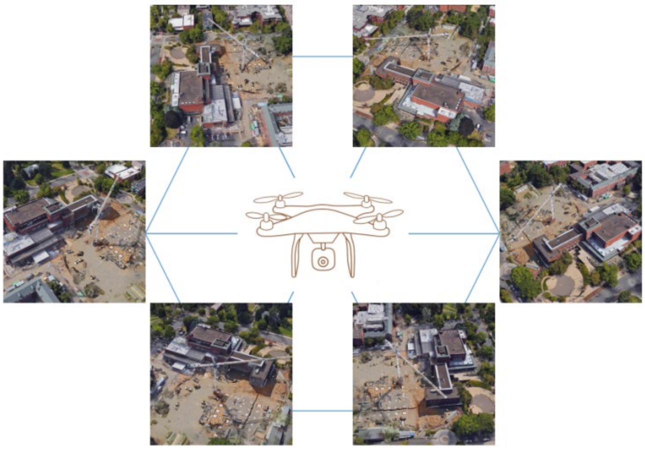

Figure 2: There are great differences in the UAV photographing the same geographical target, including proportion, viewpoint and direction.

SFM [13] is a fast and effective method to recover camera poses from multiple images with certain overlapping areas and between different perspectives. SFM obtains reliable image points with the same name through feature point detection and matching, then selects a certain number of initial seed images from all images, solves the camera parameters between seed images and recovers the positions of some encryption points, and then gradually adds associated images for reconstruction. The result after sparse reconstruction by colmap [13] is shown in Figure 3.

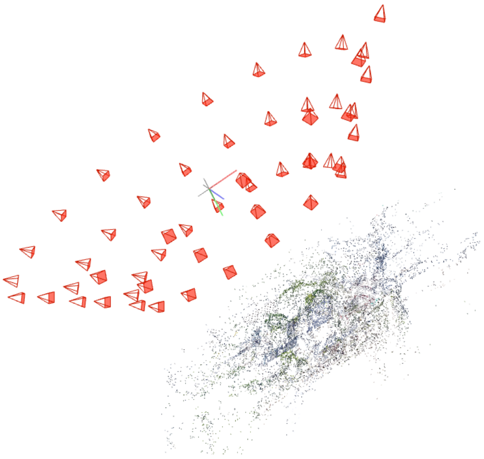

Figure 3: After sparse reconstruction by colmap, the corresponding posture information is also output.

The posture information output by colmap includes coordinates and quaternion in 3D space. The quaternion is shown in Equation (1), and its relationship with the rotation matrix is shown in Equation (2). The Angle between the two rotation matrices is calculated according to Equation (3).

$$q = w + x\vec{i} + y\vec{j} + z\vec{k} \tag{1}$$

$$\begin{bmatrix} 1-2y^2-2z^2 & 2xy+wz & 2xz-2wy \\ 2xy-2wz & 1-2x^2-2z^2 & 2yz+2wz \\ 2xz+2wy & 2yz-2wx & 1-2x^2-2y^2 \end{bmatrix} \tag{2}$$

$$2\cos(\alpha) = \mathrm{trace}(R_1^{-1}R_2) - 1 \tag{3}$$

According to the Angle obtained in Equation (3), the image is rotated. In order to ensure smooth rotation, the UAV perspective image is first cropped into a circle and then rotated. Once the orientation feature alignment is completed, the feature extraction performance of the neural network model will be easily improved.

2.2. Visual transformer network model

Liu et al. [14] proposed a SwinT network with sliding window operation and hierarchical design. SwinT network has become a new backbone network in machine vision field by using sliding window and hierarchical structure. In image classification task, with a hierarchical structure similar to that of convolution neural network to process the images, the flexible model can handle different scale images, ViT to the original network computing complexity from the index level (H×W)×2 reduced to linear level M^2×(H × W), which H represents high image, W represents the image of wide, M represents the number of patches in each window. The advantages of SwinT network are not limited by image input size, low computational complexity, and the classification performance of SwinT network in natural image datasets is also better than that of ViT network.The network structure is shown in Figure 4.

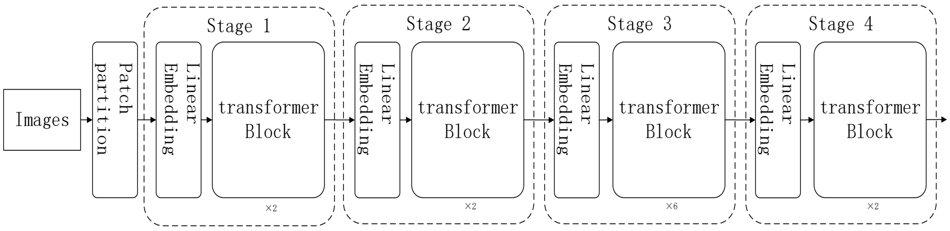

Figure 4: Network structure of swin transformer.

3. Experiment

3.1. Dataset

University-1652 dataset was used in this experiment. The dataset features images collected from three different platforms, including satellite view, UAV view and ground view. The images were collected from 1,652 buildings at 72 universities around the world. The images of the training set and the test set in the University-1652 dataset are completely independent without duplication. The training set contains 701 buildings and the test set contains 951 buildings. Each building contains one satellite view, 54 drone views, and three to four ground views. Since the problem discussed in this paper only includes drone view images and satellite view images, ground view images are not used.

3.2. Implementation Details

This research use a visual transformer with pre-trained weights to extract visual features, the training and test data are cropped and rotated on the UAV images through the orientation feature alignment strategy, and each input image is adjusted to a fixed size of 384×384 pixels. The AdamW optimization method was used to train the model, and the cross-entropy loss function was used. The learning rate was set to 0.01, the momentum was set to 0.9, and the batch size was set to 24. The experiment carried out 80 complete iterations, and the learning rate was adjusted to 0.001 when the experiment reached the 60th iteration. The model was implemented on Tensorflow and all experiments were performed on an NVIDIA RTX 3080 GPU.

3.3. Evaluation Indicators

In the experiment, Recall Rate (Recall@K) and Average Precision (AP) to evaluate the performance of the model. Recall@K Represents the proportion of correct matching images in the Ranking List of top-K, and the higher Recall@K It shows that the network performance is good. AP represents the area under the precision recall curve. The above two indicators are used as experimental evaluation criteria in UAV positioning and navigation tasks respectively.

3.4. Experiment

3.4.1. Experimental Result

We conducted ablation experiments on the model, which verified the effectiveness of our proposed method, as shown in Table 1. We compare with other advanced methods, as shown in Table 2.

Table 1: Ablation study. OA represents orientation alignment strategy, resnet50 and transformer are different neural networks backbone.

Methods	Satellite→Drone		Drone→Satellite	
	Recall@1	*AP*	Recall@1	*AP*
resnet50	74.74%	59.45%	58.23%	62.91%
transformer	75.48%	59.63%	61.17%	64.29%
OA+resnet50	77.46%	62.05%	62.76%	66.93%
OA+transformer	85.16%	69.94%	65.66%	70.33%

Table 2: Comparison with other advanced methods.

Methods	Satellite→Drone		Drone→Satellite	
	Recall@1	*AP*	Recall@1	*AP*
Baseline [5]	74.74%	59.45%	58.23%	62.91%
LCM [15]	79.89%	65.38%	66.65%	70.82%
SFPN [16]	80.26%	71.58%	70.83%	77.36%
This research	85.16%	69.94%	65.66%	70.33%

3.4.2. Retrieval result display

As shown in Figure 5, the results of the image retrieval experiment, the upper part is the UAV image retrieval satellite image, and the lower part is the satellite image retrieval UAV image.

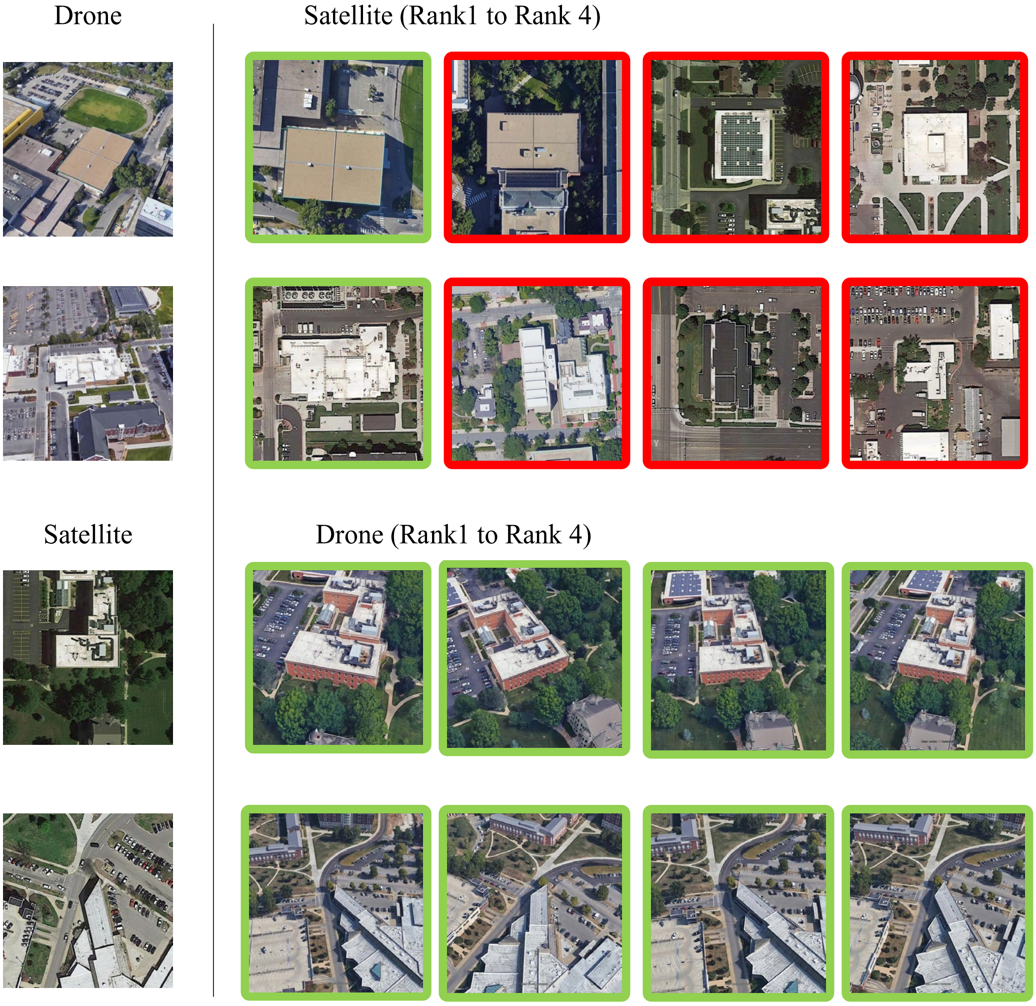

Figure 5: Use the original image as an effect display. The search results are displayed with green boxes representing correct and red boxes representing wrong.

4. Conclusions

The SFM algorithm combined with the visual transformer model is used for cross-view image matching. It makes full use of UAV pose information and orientation features to learn the shared space information of satellite view image and UAV image. At the same time, the network model combines images from different perspectives into the same branch, which effectively solves the problem of data sample size imbalance and strengthens the learning of similar features of cross-view images.

The next step of this research will consider further using the pose information of UAV, such as coordinate information, to train an adaptive view conversion method for spatial domain difference through different height differences and different visual field ranges.

References

[1] Vo, N., Jacobs, N. and Hays, J. (2017). Revisiting im2gps in the deep learning era. In Proceedings of the IEEE international conference on computer vision (pp. 2621-2630).

[2] Toker, A., Zhou, Q., Maximov, M. and Leal-Taixé, L. (2021). Coming down to earth: Satellite-to-street view synthesis for geo-localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6488-6497).

[3] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[4] Bay, H., Tuytelaars, T. and Gool, L. V. (2006). Surf: Speeded up robust features. In European conference on computer vision (pp. 404-417). Springer, Berlin, Heidelberg.

[5] Zheng, Z., Wei, Y. and Yang, Y. (2020). University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In Proceedings of the 28th ACM international conference on Multimedia (pp. 1395-1403).

[6] Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B. and Yang, Y. (2021). Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology, 32(2), 867-879.

[7] Hu, S. and Chang, X. (2020). Multi-view drone-based geo-localization via style and spatial alignment. arXiv preprint arXiv:2006.13681.

[8] Tian, Y., Chen, C. and Shah, M. (2017). Cross-view image matching for geo-localization in urban environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3608-3616).

[9] Hu, S., Feng, M., Nguyen, R. M. and Lee, G. H. (2018). Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7258-7267).

[10] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5297-5307).

[11] Regmi, K. and Shah, M. (2019). Bridging the domain gap for ground-to-aerial image matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 470-479).

[12] Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J. and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8798-8807).

[13] Schonberger, J. L. and Frahm, J. M. (2016). Structure-from-motion revisited. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4104-4113).

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

[15] Ding, L., Zhou, J., Meng, L. and Long, Z. (2020). A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. Remote Sensing, 13(1), 47.

[16] He, S. and Wang, Y. (2021). Cross-view geo-localization via Salient Feature Partition Network. In Journal of Physics: Conference Series (Vol. 1914, No. 1, p. 012009). IOP Publishing.