Cross-view Image Geo-Localization using Multi-Scale Generalized Pooling with Attention Mechanism

Duc Viet Bui[*]
Department of Computer Science, National Defense Academy, 1-10-20 Hashirimizu
Yokosuka City, Kanagawa Prefecture, Japan[†]

Masao Kubo
Department of Computer Science, National Defense Academy, 1-10-20 Hashirimizu
Yokosuka City, Kanagawa Prefecture, Japan

Hiroshi Sato
Department of Computer Science, National Defense Academy, 1-10-20 Hashirimizu
Yokosuka City, Kanagawa Prefecture, Japan
E-mail: ed21009@nda.ac.jp, masaok@nda.ac.jp, hsato@nda.ac.jp

Abstract

Cross-view image matching for geo-localization is the task of finding images containing the same geographic target across different platforms. This task has drawn significant attention due to its vast applications in UAV's self-localization and navigation. Given a query image from UAV-view, a matching model can find the same geo-referenced satellite image from the database, which can be used later to precisely locate the UAV's current position. Many studies have achieved high accuracy on existing datasets, but they can be further improved by combining different feature processing methods. Inspired by previous studies, in this paper, we proposed a new strategy by using a channel-based attention mechanism with a generalized mean pooling method to enhance the feature extracting process, which improved accuracy.

Keywords: Cross-view image matching, UAV, attention mechanism, generalized mean pooling.

1. Introduction

The applications of unmanned aerial vehicles (UAV) in daily life have been rapidly increasing. UAV has become an essential part of various fields such as aerial surveillance[1], agriculture[2], transportation, search and rescue missions. Along with their applications, to further reduce human's work, the need for autonomous drones has been increasing over time. However, most researches failed to achieve a fully autonomous drone system, as the most used navigation system (Global Positioning System - GPS) has many limitations in real-life missions. For example, GPS is not powerful enough when high buildings or trees block GPS signals, leading to difficulties in applying UAVs in cities and urban areas.

Many solutions for navigation in autonomous drone systems have been proposed, and among them, cross-view image matching-based methods have received lots of researchers' attention due to the vast application value in geo-localization[3,4,5,6)]. Cross-view image matching is the task of matching a satellite-view image with

.

geographic location tags and a UAV-view image without a geographic location tag or vice versa to locate a UAV's position based on information from taken images. Figure 1 shows an example of cross-view matching methods. Given a UAV-view image of a building, the matching model searches for the image of that building in the satellite-view image gallery. The output is a satellite-view image similar to the query UAV-view image. This output can be used to locate the current position of the UAV.

Fig. 1: Example of UAV-view → Satellite-view

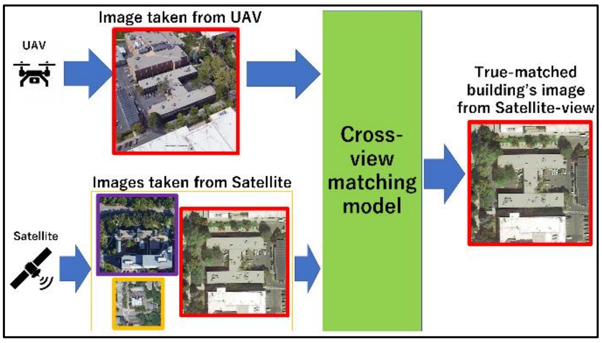
Image taken from UAV
UAV
Cross-view matching model
True-matched building's image from Satellite-view
Images taken from Satellite
Satellite

Early-stage cross-view image matching researches[7] focused on using traditional image processing methods, which used hand-crafted features from images. In recent years, with the rapid development of machine learning and deep learning in image processing, especially the convolutional neural network, many studies attempted to apply them in cross-view image matching for geo-localization, some of which have achieved significant results[3,6)]. Also, deep learning's well-known self-attention mechanisms have been used to understand image representations further and bridge the gap between images from different views[8]. However, the self-attention mechanisms usually require lots of computational costs, which is entirely unsuitable for current UAV systems. Moreover, pooling strategies in the previous cross-view matching method were mainly average pooling and max pooling, limiting the feature learning process from extracting important global features.

Inspired by cross-view image matching related works, in this paper, we ensembled several feature processing methods with a channel-based attention mechanism and multi-scale generalized pooling strategy. Our proposed model has shown an increase in performance through experiments compared to the state-of-the-art method (SOTA).

2. Related works

The previous studies[9,10)] often consider cross-view image matching as an image retrieval problem since they aim to find similar images to a query image among an image dataset. In the cross-view matching problem, the main task is to learn image representation that varies in different views, thus bridging the gap between multi-view images. The schematic diagram of cross-view matching is described in Figure 2.

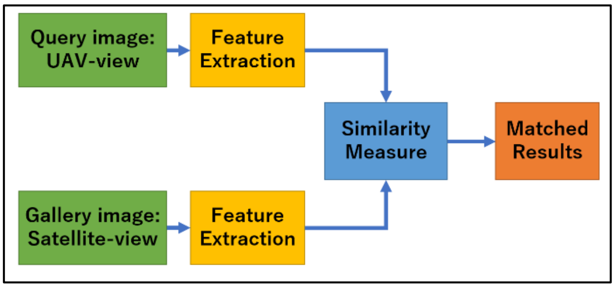

Fig. 2: The schematic diagram of cross-view image matching problem (UAV-view → Satellite-view)

At first, features from query images and a database (gallery images) are extracted using different feature processing methods. After that, features' similarities were calculated by using distance similarity measures such as cosine similarity or Euclidean distance. The results are later used to create a ranking list, from which the model will determine the true-matched image to the query image.

Traditional image processing methods such as SIFT[11] or SUFT[12] have been implemented in early research for the feature extracting phase. However, as the gap between different viewpoints is enormous, the matching results are not high as expected. As a result, more and more researchers have been paying attention to the powerful convolutional neural networks (CNN), which is well-known for its abilities to learn high-level features. Another work focused on learning discriminative features using metric learning and proposing various loss functions. This line of work has similar approaches with face verification and person re-id problems cause they adopt ranking losses such as contrastive loss or triplet loss to learn the relative distances between different inputs. For example, Lin et al.[13] adopted contrastive loss to optimize network parameters, and Hu et al.[14] designed

CVM-Net, which employed weighted-soft-margin ranking loss.

Moreover, some works try to enhance the feature learning phase by applying the attention mechanism. Attention-mechanism is a method invented to make neural networks learn the most relevant features from inputs, thus increasing the network's learning abilities. Novel self-attention model – Transformer[15] has been studied in many natural language processing problems, and now its application in vision processing (known as Vision Transformer[16]) has been adopted in numerous cross-view matching researches[8]. However, the disadvantages of Transformer architecture are high computational cost and a massive amount of data required for training.

Generalized mean pooling (GeM pooling) was first proposed in Ref. 17 as an alternative pooling method for image retrieval. Since then, it has been widely applied in many retrieval systems and achieved promising results.

Following related works on cross-view image matching, the architecture of our proposed model was based on Siamese networks (twin neural network), with two branches for each view's input, and we investigated the effectiveness of the channel-based attention with GeM pooling towards the cross-view matching problem.

3. Materials and the proposed method

3.1. Dataset and evaluation metrics

In this work, we use the University-1652 dataset published by Zheng et al.[18], as it is the only benchmark dataset with both satellite-view and UAV-view images, which helps solve cross-view matching for UAV navigation. This dataset contains 1652 geographic targets from 72 universities all over the world. Each target contains three views: satellite-view, UAV-view, and street-view. To reduce the high cost in airspace control and flying UAV, all UAV-view and street-view images were collected by a 3D engine named Google Earth, while satellite-view images were captured by Google Map. All images in the dataset have geo-tags as their class labels. The view of UAVs in Google Earth was controlled by simulated camera-view, and the height of view descends from 256 to 121.5m. Each target consisted of 1 satellite-view image, 54 UAV-view images, and a few street-view images. The dataset was split into the training and test sets with no overlapped classes. The captured images have an original size of 512x512. The distribution of data in each set is described in Table 1.

Table 1. Distribution of data in University-1652

	Images	**Class**	**Universities**
Training	50218	701	33
Query (UAV)	37855	701	39
Query (Satellite)	701	701	
Query (Ground)	2579	701	
Gallery (UAV)	51355	701	
Gallery (Satellite)	951	951	
Gallery (Ground)	2921	793	

Regarding the evaluation metrics, most of the image retrieval and cross-view image matching researches has been using Recall@K and Average Precision (AP) as the main indicator for evaluating proposed systems. Recall@K is computed by calculating the ratio of the true-matched image in the top-K results of the ranking list. On the other hand, AP is a popular metric in measuring the precision of a retrieval system. The higher Recall@K and AP, the better the model performs.

3.2. Proposed method

The overview of proposed network architecture is described in Figure 3. From subsection 3.2.1 to 3.2.3, we explain the major components of the model.

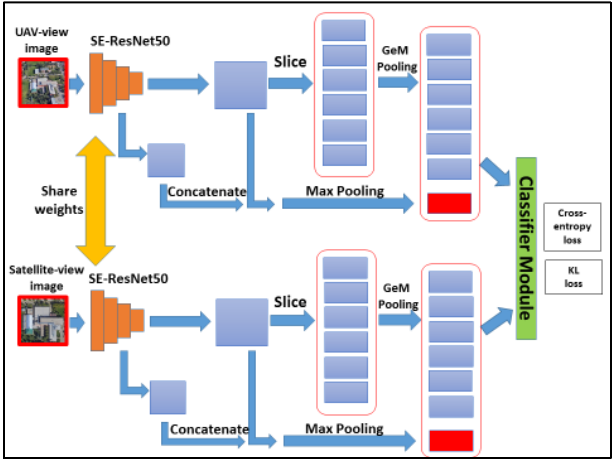

Fig. 3: Proposed network architecture

3.2.1. *Feature extraction strategy*

Because of excellent accuracy and inference time, other existing methods used backbone from ResNet50[19] model or VGG16[20] as the main feature extractor, while the usage of attention modules in these backbones is rarely seen. However, we believe that an attention mechanism can strengthen the saliency value of each view and restrain the unnecessary features from affecting the final results. Therefore, among various attention mechanisms in literature, channel-based attention - the SE block from Ref.21 is chosen for its efficiency in reinforcing the backbone while requiring very little computation cost. The SE block can also be easily implemented in ResNet50's Residual block as follow (Figure 4).

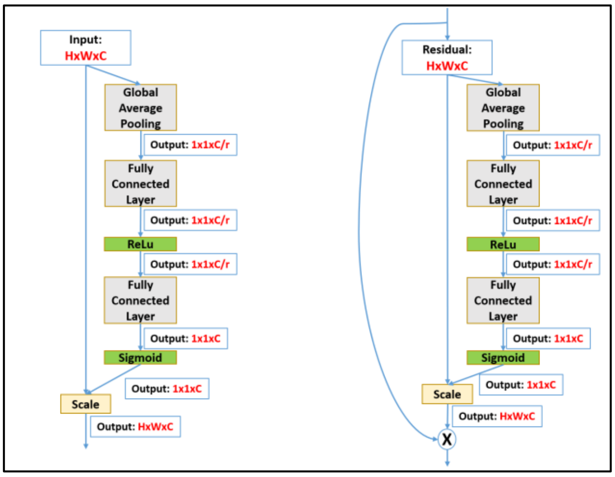

Fig. 4: SE block (left) and its implementation in Residual block (right)

Deep layers in CNN tend to learn high-level features (object, human), while shallow layers extract low-level features such as shapes, edges in the image. Geographic targets in the dataset are mainly buildings, so these low-level features play an important role in understanding the entire view. For that reason, in our proposed model, not only the last layer of the model but the global features from shallow layers (here, we chose the third and the fourth layers) are also extracted. Feature maps were concatenated, and Global max pooling method was applied after that.

3.2.2. *Multi-scale block and pooling strategy*

Previous works in Ref. 22 and Ref. 23 proposed a feature partition strategy to take advantage of contextual information. In particular, the output feature map is divided into several blocks, and then the global average pooling method is performed. Here we also applied the block strategy, but GeM pooling was put in practice instead of global average pooling. The formula of GeM pooling can be defined as follow:

$$f^{(g)} = [f_1^{(g)}..f_k^{(g)}..f_K^{(g)}]^T, f_k^{(g)} = \left(\frac{1}{|X_k|}\sum_{x \epsilon X_k} x^{p_k}\right)^{\frac{1}{p_k}} \qquad (1)$$

with X_k represents feature map, K is the number of channel and p_k is the pooling parameter. This pooling parameter can be manually set or changed through learning process.

The way of creating multi-scale feature blocks is described in Figure 5:

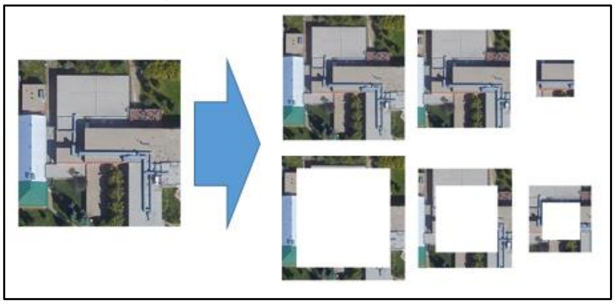

Fig. 5: Multi-scale block partition

3.2.3. *Loss functions*

In Ref. 18, instance loss was adopted to train the multi-branches networks, and results have shown the good effect of this loss function compared to other ranking losses in terms of cross-view matching accuracy. Instance loss was first proposed in Ref. 24, an alternative way to learn the distance between features. Extracted features from each branch were sent to the shared fully connected layer in order to map the features of all sources into one shared feature space. Finally, the cross-entropy loss function was applied to optimize the network.

Additionally, in Ref. 23, Kullback-Leibler divergence (KL divergence) was first applied in the training phase. In the fields of machine learning, KL divergence is a measure of how a probability distribution differs from another probability distribution. KL divergence is expected to close the gap between two different domains (UAV and satellite).

Here we use the Softmax function to obtain the normalized probability scores, and KL divergence is then computed and added to the training loss. The KL divergence formula is defined as follow:

$$L_{KL(p_2||p_1)} = \sum_{n=1}^{N} p_2^n log \frac{p_2^n}{p_1^n} \qquad (2)$$

p_1, p_2 are predicted results of each branch.

4. Experiments and discussions

4.1. Experiments

4.1.1. *Implementation details*

We performed training our proposed model with the University-1652 dataset and some ablation experiments with different backbones and pooling layers to understand the effect of attention mechanism and pooling strategy towards the results.

The experiments were carried out with three different input sizes (256x256, 384x384, 512x512). Random flipping and random cropping were used to augment the training images. SE-ResNet50 was pre-trained with the ImageNet dataset, and the stride of the final down-sampling layer was fine-tuned from 2 to 1 to increase the size of the feature map output by the backbone. The optimizer was Stochastic Gradient Descend (SGD) with a momentum of 0.9 and weight decay of 5x10^{-4}. The training lasted for 120 epochs, with an initial learning rate of 1x10^{-4} for backbone layers and 1x10^{-3} for other layers. The pooling parameter p_k in GeM pooling was initially set to 3.

The classifier layer was removed in the testing phase, so the model returns only extracted features as outputs. Euclidean distance was applied to compute the similarity between feature vectors from different views.

4.1.2. *Experiment Results*

In Table 2, we compared our proposed method with the SOTA in Ref. 23. In Table 3, we showed the results of the ablation experiments on SE block and pooling strategy.

Table. 2 Comparison of proposed method with SOTA

Method	UAV → Satellite		Satellite → UAV	
	R@1	**AP**	**R@1**	**AP**
MSBA (256x256)	82.33	84.78	90.58	81.61
MSBA (384x384)	86.61	88.55	92.15	84.45
MSBA (512x512)	86.69	88.66	92.01	84.45
Ours (256x256)	**82.87**	**85.13**	**90.87**	**82.06**
Ours (384x384)	**86.96**	**88.88**	**92.30**	**84.92**
Ours (512x512)	**87.90**	**89.71**	**92.58**	**85.49**

Table. 3 Results of ablation experiments of using different strategies

Method	UAV → Satellite		Satellite → UAV	
	R@1	**AP**	**R@1**	**AP**
ResNet50 + Average Pooling (256x256)	78.99	81.85	87.30	78.18
ResNet50 + GeM Pooling (256x256)	79.58	82.22	86.16	78.56
SE-ResNet50 + Average Pooling (256x256)	81.44	84.09	90.16	79.44
SE-ResNet50 + GeM Pooling (256x256)	82.87	85.31	90.87	82.06

4.2. *Discussions*

From the results of the experiments shown in Table 2, it can be seen that our proposed model has an increase in terms of accuracy. The improvement in 256x256 and 384x384 input size was little, but in the original size (512x512), we achieved a slight increase compared to SOTA methods (1% or more in all evaluation indicators). Table 3 gave us a more specific view of how SE block and GeM Pooling affected the final results. GeM pooling showed its good performance in cross-view matching problems when using the same ResNet50 backbone, and the attention feature created by the SE block gave an important contribution to the final learning results compared to the basic residual block (approximately 4% increase in accuracy). This gave promising capability of doing further studies on the effectiveness of SE block and GeM pooling on the cross-view problem.

5. Conclusion

In this paper, we addressed the problems in navigating UAVs without GPS and focused on solving cross-view image matching tasks for navigation. We designed a network using a channel-based attention mechanism with a multi-scale generalized mean pooling strategy and verified its effectiveness on the University-1652 dataset. Experiment results showed that our proposed model has an increase in accuracy compared to the previous SOTA result. In future works, to increase the model's robustness to features in cross-view domains, we plan to further utilize the attention mechanism in the feature extracting phase, and some other pooling strategies are under-considered. Instead of 3D images, real environment data should also be tested in the next phase of our research.

References

1. M. Kontitsis, K. P. Valavanis and N. Tsourveloudis: "A UAV vision system for airborne surveillance", IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, New Orleans, LA, USA, pp. 77-83 Vol.1, 2004,
2. S. W. Chen, et al: "Counting apples and oranges with deep learning: a data-driven approach", IEEE Robotics and Automation Letters, vol. 2, no. 2, 2017,
3. L. Ding, et al.: "A Practical Cross-View Image Matching Method between UAV and Satellite for UAV-Based Geo-Localization.", Remote Sensing 13.1: 47, 2021,
4. L. Liu, H. Li, and Y. Dai.: "Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization", Proceedings of the IEEE International Conference on Computer Vision, pp. 2570–2579, 2019.
5. Y. Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. : "Optimal Feature Transport for Cross-View Image Geo-Localization". arXiv preprint arXiv:1907.05021 , 2019.
6. Y. Tian, C. Chen, and Mubarak Shah: "Cross-view image matching for geo- localization in urban environments", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3608–3616, 2017,
7. F. Castaldo et al.: "Semantic cross-view matching", Proceedings of the IEEE International Conference on Computer Vision Workshops, p. 9-17, 2015.
8. H. Yang et al.: "Cross-view Geo-localization with Evolving Transformer",arXiv preprint arXiv:2107.00842, 2021.
9. N. Khurshid et al.: "Ground to Aerial Image Retrieval Through Deep Learning", Proceeding of International Conference on Neural Information Processing, Springer, p. 210-221, 2019.
10. S. Zhu et al.: "Cross-view image geo-localization beyond one-to-one retrieval", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 3640-3649, 2021.
11. D. G. Lowe et al.: "Object recognition from local scale-invariant features", Proceedings of the IEEE International Conference on Computer Vision, 1999.
12. H. Bay, T. Tuytelaars, and L. Van Gool.: "Surf: Speeded-up robust features", Proceedings of European Conference on Computer Vision, 2006.
13. T. Lin et al.: "Learning deep representations for ground-to-aerial geolocalization.", Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
14. S. Hu et al.: "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7258–7267, 2018.
15. V. Ashish, et al.: "Attention is all you need", Advances in neural information processing systems. pp. 5998-6008, 2017.
16. A. Kolesnikov et al. "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv preprintarXiv:2010.11929,2020.
17. F. Radenović et al.: "Fine-tuning CNN image retrieval with no human annotation", IEEE transactions on pattern analysis and machine intelligence, 41(7), pp 1655-1668, 2018.
18. Z. Zheng, Y. Wei, and Y. Yang. : "University-1652: A multi-view multi-source benchmark for UAV-based geo-localization.", Proceedings of the 28th ACM international conference on Multimedia, 2020.
19. H. Kaiming, et al. : "Deep residual learning for image recognition.", Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
20. K. Simonyan, Z. Andrew. : "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv 1409.1556, 2014.
21. J. Hu, L. Shen, and G. Sun.: "Squeeze-and-excitation networks.", Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
22. W. Tingyu, et al.: "Each part matters: Local patterns facilitate cross-view geo-localization.", IEEE Transactions on Circuits and Systems for Video Technology, 2021.
23. J. Zhuang et al.: "A Faster and More Effective Cross-View Matching Method of UAV and Satellite Images for UAV Geolocalization", Remote Sensing, 13.19: 3979, 2021,
24. Z. Zheng, et al. "Dual-Path Convolutional Image-Text Embeddings with Instance Loss." arXiv preprint arXiv:1711.05535,2017.

Authors Introduction

Mr. Duc Viet Bui

He received his M.S degrees from Department of Computer Science, National Defense Academy of Japan in 2021. He is currently a doctoral student at Department of Computer Science in National Defense Academy of Japan. His research related to different applications of computer vision in aerial robotics. His interests include computer vision, machine learning, deep neural networks and aerial robotics.

Dr. Masao Kubo

He is Associate Professor of Department of Computer Science at National Defense Academy in Japan. He graduated from the precision engineering department, Hokkaido University, in 1991. He received his Ph.D. degree in Computer Science from the Hokkaido University in 1996 (multi-agent system). He had been the research assistant of the chaotic engineering Lab, Hokkaido university. He was a visiting research fellow of Intelligent Autonomous Lab, university of the west of England (2005). He is the associate professor of information system lab, Dep. of Computer Science, National Defense Academy, Japan. His research interest is Multi agent system.

Dr. Hiroshi Sato

He is an Associate Professor of the Department of Computer Science at the National Defense Academy in Japan. He obtained a degree in Physics from Keio University in Japan and degrees of Master and Doctor of Engineering from Tokyo Institute of Technology in Japan. He was previously a Research Associate at the Department of Mathematics and Information Sciences at Osaka Prefecture University in Japan. His research interests include agent-based simulation, evolutionary computation, and artificial intelligence. Dr. Sato is a member of the Japanese Society for Artificial Intelligence (JSAI), Society of Instrument and Control Engineers (SICE), and The Institute of Electronics, Information and Communication Engineers. IEICE). He was the editor of IEICE and SICE.