Efficient UAV localization using combined autoencoder and SIFT

Ngo Van Quan[1], Phan Huy Anh[2*], Bui Thi Thanh Tam[2], Nguyen Chi Thanh[1]

[1]Institute of Information Technology, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam;
[2]Institute of Electronics, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam.
[*]Corresponding author: huyanhfanvdt@gmail.com
Received: 04 Sep. 2024; Revised 11 Nov. 2024; Accepted 15 Nov. 2024; Published 06 Dec. 2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.FEE.2024.142-148

ABSTRACT

In GNSS-denied environments, accurate Unmanned Aerial (UAV) localization faces significant challenges. This paper introduces a vision-based localization method combining autoencoder and SIFT algorithms, referred to as AE+SIFT. The method compresses high-resolution map images into low-dimensional vectors, which are stored in a database for efficient retrieval. During the localization process, UAV images are encoded and matched with the database, followed by SIFT and homography projection for precise positioning. The AE+SIFT approach enhances localization accuracy, achieving an average coordinate error of 3.94 meters relative to the ground truth. Notably, when UAV images are misaligned with reference images, our method outperforms the existing AE method in terms of accuracy.

Keywords: GNSS; Autoencoder; SIFT; Visual Localization; UAV.

1. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are increasingly essential in fields like surveillance, precision agriculture, and disaster management [1, 2]. However, reliable localization in GNSS-denied environments remains challenging [3]. Traditional GNSS-based methods are prone to interference and signal issues [4], prompting interest in vision-based techniques using aerial imagery [5].

Vision-based localization matches UAV-captured images with reference maps, relying on feature extraction from satellite images or smaller reference images [6]. While accurate, these methods often involve high computational costs and large datasets [7]. Additionally, inefficiencies in feature extraction and image registration can reduce accuracy [8, 9].

Various approaches address GNSS-denied localization [10-15]. For example, [16] employs ORB, a local feature extractor [7], while [17] uses neighbor consensus networks to detect patterns in dense feature correspondences. [18] and [19] enhance feature matching with LightGlue and LoFTR, integrating global context via Transformers. These methods involve matching UAV images against large reference datasets, leading to slow inference times.

A more efficient two-stage approach narrows candidate images before detailed matching. Autoencoders, as in [20] and [21], compress images into low-dimensional vectors for faster retrieval. Transformer-based models [18] further improve encoding and registration but still rely on primitive matching techniques.

To address these issues, we propose combining autoencoder and SIFT for UAV localization using a custom TIFF map. Overlapping grid images are compressed into vectors via autoencoder [13] and stored in a database for efficient retrieval [10]. Captured UAV images are encoded similarly, compared to the database, and matched using SIFT and homography projection for precise geolocation [11].

The paper is structured as follows: Section 2 reviews related work, section 3 outlines our methodology, section 4 presents experimental results, and section 5 concludes with future directions.

2. PROBLEM STATEMENT AND METHODOLOGY

In this section, we propose an image-based localization method that combines autoencoder and SIFT techniques to enhance accuracy, making it viable for real-world flight scenarios without GPS signals. The method includes two phases: an offline phase for data preparation and model training, and an online phase for map image registration using the AE encoder and localization via SIFT. Figure 1 illustrates the overall methodology.

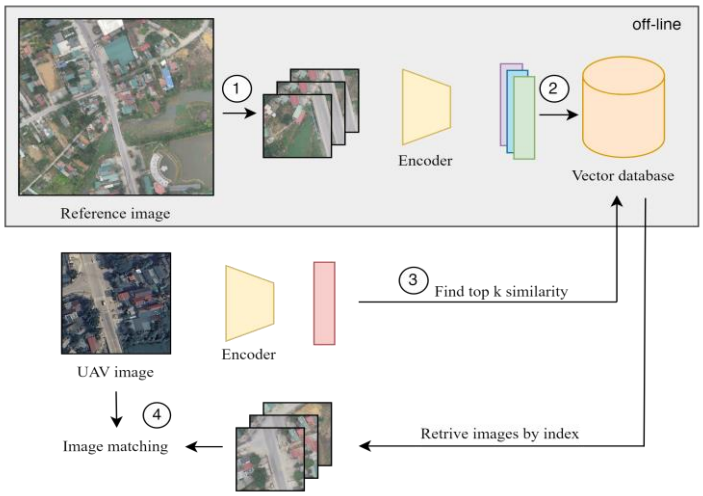

Figure 1. *Autoencoder+SIFT for UAV localization: (1) The reference image cropped into smaller patches and encoded for feature vectors. (2) Store the vectors in a database. (3) Encode and compare UAV image with the database to find the most similar candidates. (4) Matching UAV with image candidates using SIFT.*

2.1. Data Preparation and Training Autoencoder

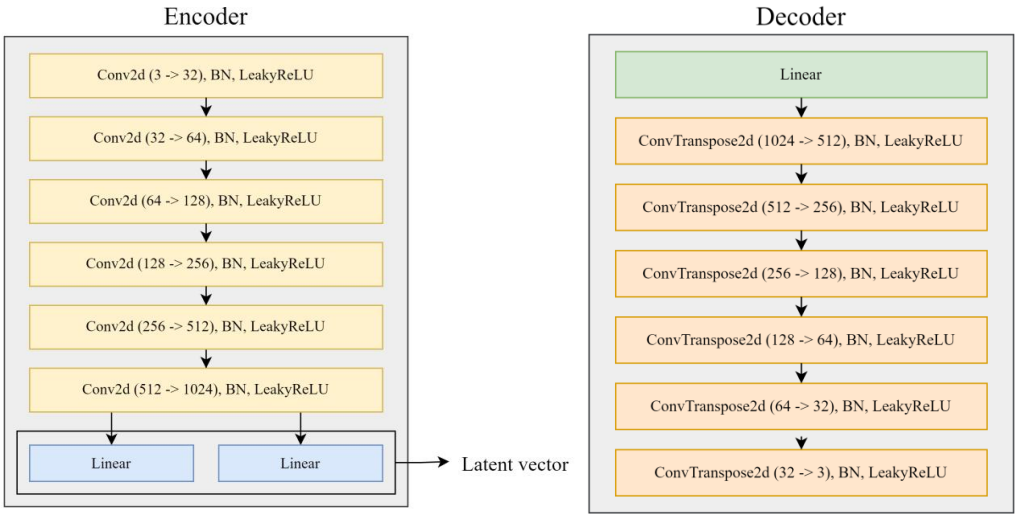

Figure 2. *The autoencoder architecture.*

The map and reference image datasets are cropped and overlapped from the .tif map file into square pixel-based images. The map images encompass a larger area than the reference images, ensuring the UAV flight path is fully contained within the reference images. The autoencoder comprises an encoder for extracting features into low-dimensional vectors and a decoder for reconstructing the original image. Inspired by [22], the architecture employs a loss function combining Mean Squared Error

(MSE) and Kullback-Leibler Divergence (KLD), as shown in figure 2.

Specifically, the encoder compresses the input image using multiple 2D convolutional (Conv2d) layers, each followed by batch normalization (BN) and LeakyReLU activations. As the number of feature maps increases progressively, the spatial dimensions are reduced. The encoder's final output is a latent vector produced through fully connected (Linear) layers. The decoder reconstructs the input from the latent vector using transposed convolutional (ConvTranspose2d) layers, following the same pattern of batch normalization and LeakyReLU activations. It progressively increases the spatial dimensions to restore the original input size. The loss function is calculated as follows [22]:

$$L = L_{MSE} + kld_{weight} \times L_{KLD} \tag{1}$$

where,

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_i - \hat{x}_i^{\ 2} \tag{2}$$

$$L_{KLD} = -\frac{1}{2}\ 1 + \log\ \sigma^2\ - \mu^2 - \sigma^2 \tag{3}$$

where N is the number of image dimensions, $\hat{x}_i$, x_i is the input and reconstructed output values at position i, respectively, μ is the mean of the latent distribution, $\log\ \sigma^2$ is the log variance of the latent distribution, kld_{weight} helps balance data reconstruction with aligning the latent space to the standard distribution.

The autoencoder is trained on a large dataset containing over 20,000 randomly clipped map images to ensure effective feature extraction. At the conclusion of the training process, the autoencoder's parameters are saved. The trained encoder then encodes the reference image set into a collection of feature vectors called a vector database.

2.2. Reference Image Registration with the Encoder and Localization Using SIFT

The UAV image is encoded into a feature vector using the encoder from the trained autoencoder. Subsequently, the image registration process involves searching for the most similar reference images in the vector database generated during the offline phase, using this feature vector, and returning a list of images with characteristics most closely matching the UAV image. The reference images are ranked according to their similarity, and their indices are used to determine the potential positions of the UAV on the map.

Then, the SIFT algorithm is applied to detect and match keypoints between the UAV image and the reference images. The keypoints are used to compute the homography matrix, which transforms the UAV image into the reference image's coordinate system. Once the homography matrix is calculated, the corners of the UAV image are transformed to determine its position on the map. Next, the keypoints are used to determine the central coordinates of the UAV within the reference image's coordinate system. The UAV's latitude and longitude are then calculated based on the corners of the image aligned with the map. If the number of matched keypoints is adequate and the homography calculation is accurate, the UAV's position can be determined with high precision. The UAV's latitude and longitude coordinates are then calculated using the following formula:

$$\begin{aligned} Lat &= Lat_{tl} + \frac{C_y}{H} \times\ Lat_{br} - Lat_{tl} \\ Lon &= Lon_{tl} + \frac{C_x}{W} \times\ Lon_{br} - Lon_{tl} \end{aligned} \tag{4}$$

where W and H are the reference image's width and height. Lat_{tl}, Lon_{tl}, Lat_{br}, and Lon_{br} are the top-left and bottom-right geographical coordinates of the reference image, respectively. C_x and C_y are the pixel coordinates of the reference image's center. Lat and Lon are the calculated coordinates for the UAV image.

3. RESULTS AND DISCUSSION

The performance of the proposed localization method (AE+SIFT) is compared to that of SIFT-based and AE-based localization methods on the same dataset. Assuming we know in advance that the UAV's flight area is within the reference map region and that the UAV is flying at a constant altitude with its images aligned with the reference images. We evaluate the proposed solution in two scenarios: first, when the UAV image is rotated to match the angle of the reference image, and second, when the UAV image is not aligned with the reference image angle.

The map image dataset, consisting of 20,000 images with a resolution of 1500x1500 pixels, is divided into training and test sets with an 80/20 split ratio. The cropped images from the reference image are utilized as the reference image set during the evaluation phase. The UAV flight video in the .mp4 file is 50 seconds long, recorded at 30 frames per second, and covers a distance of approximately 600 meters. The autoencoder was trained for 200 epochs, using a batch size of 64, a learning rate of 0.0001 and kld_{weight} = 0.00025.

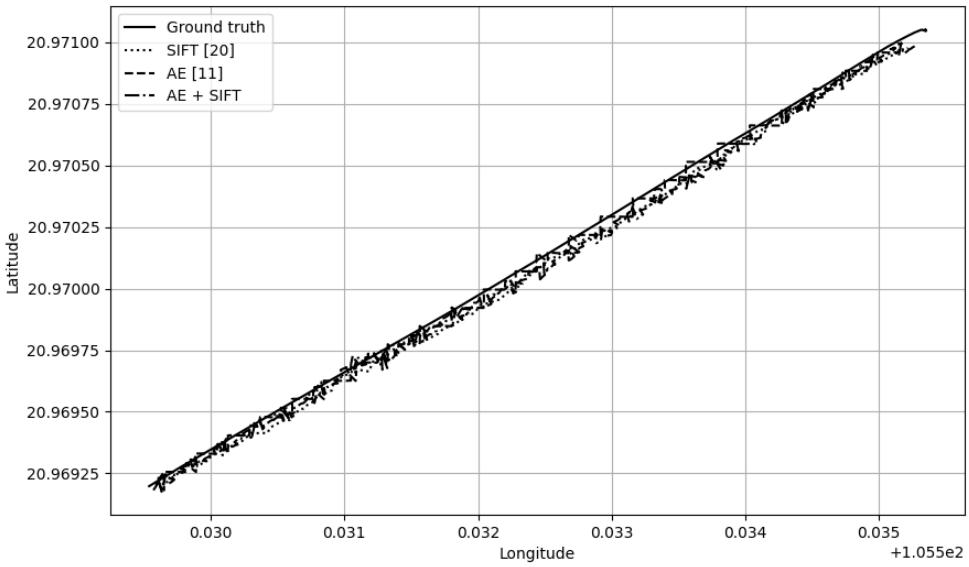

Figure 3. *The deviation of the positioning methods from the ground-truth path.*

We first assess the impact of the reference image dimensions on positioning accuracy when the UAV images are aligned with the reference images. Table 1 presents a comparison between our method (AE+SIFT) and the AE method [20] based on the average coordinate error between the UAV and ground truth across different reference image sizes. Table 1 indicates that for smaller square image sizes (2900 to 3500 pixels), the AE method achieves lower errors, particularly with an error of 2.31 m at 2900 pixels, making it more suitable for these resolutions. However, at the largest image size (4000 pixels), AE's error sharply rises to 91.8 m, highlighting its limitations with larger images. In contrast, AE+SIFT demonstrates more consistent performance across all sizes, performing especially well at 4000 pixels with a significantly lower error of 8.9 m despite showing slightly higher errors than AE at smaller resolutions.

Table 1. *The average coordinate error comparison in meters for methods across various image sizes.*

Image sizes / Methods	2900	3100	3300	3500	4000
AE [20]	2.31	2.6	2.54	8.5	91.8
AE+SIFT	4.4	4.09	3.94	4	8.9

Figure 3 shows the deviation of the positioning methods from the ground-truth path. The graph shows that all three methods (SIFT, AE, and AE+SIFT) align closely with the ground truth, indicating strong overall accuracy in path estimation. Of the three, the AE+SIFT method (dash dot) aligns most closely with the ground truth, effectively minimizing deviations along the trajectory. While SIFT (dot) and AE (dash) methods also perform well, AE alone exhibits slightly larger deviations in some sections, suggesting that the AE+SIFT method enhances positional accuracy.

In real-flight conditions, UAV-collected images may not be aligned with the reference images, which can impact the accuracy of coordinate determination for positioning methods. Therefore, in the following section, we assess the positioning accuracy when UAV images are not aligned with the reference images. Figure 4 illustrates the performance of the AE+SIFT method with UAV data in two scenarios: with rotation (a) and without rotation (b) relative to the reference images. In the rotated case (a), the UAV images (red squares) show better alignment with the image, and the green keypoint matches are denser and more evenly distributed. Conversely, without rotation (b), the UAV images exhibit less accurate alignment, and the keypoint matches are more scattered, suggesting that rotating the UAV images enhances both alignment and keypoint matching performance.

Figure 4. *UAV Image Alignment: Rotated vs. wo/ Rotated with Reference image.*

Table 2. *Average coordinate error (meters) for different methods in UAV-image misalignment case.*

Image sizes / Methods	2900	3100	3300	3500
AE wo/rotate [20]	109.82	100.56	101.24	116.81
AE+SIFT wo/rotate	*21.7*	*23.05*	*28.3*	*34.94*

Table 2 indicates that AE without rotation exhibits significantly higher coordinate errors across all image sizes, with values ranging from 100.56 m to 116.81 m, highlighting its poor performance, especially at larger image sizes such as 3500 pixels. In contrast, AE+SIFT without rotation consistently performs better, with much lower errors (21.7 m to 34.94 m), demonstrating greater accuracy and robustness across various image sizes, even when rotation is not applied.

4. CONCLUSIONS AND FUTURE WORKS

The AE+SIFT method significantly improves UAV localization accuracy compared to AE alone, particularly in cases of image misalignment, by maintaining lower coordinate errors across varying image sizes. Although the proposed method is robust in different flight conditions, its performance declines in non-rotated cases, indicating a need for further improvements in addressing severe misalignments or more complex environments. Future research could aim to enhance the autoencoder's generalization capabilities across diverse environments and optimize the image registration process to boost speed and adaptability for real-time UAV applications in more extensive and more challenging areas. Moreover, the proposed algorithm should be deployed and tested on embedded systems together with the adoption of multi-source data fusion techniques for real-life UAV applications.

REFERENCES

[1]. Yang, Z., et al., *"A Survey on UAV-Based Applications for Smart Agriculture,"* IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4132-4149, (2021).

[2]. Li, Y., et al., *"Precision Agriculture with UAV-Based Remote Sensing: A Review,"* Remote Sensing, vol. 12, no. 9, pp. 1-22, (2020).

[3]. Deng, X., et al., *"A Robust Vision-Based Localization System for UAVs in GPS-Denied Environments,"* IEEE Transactions on Instrumentation and Measurement, vol. 70, (2021).

[4]. Humphreys, T. E., et al., *"Assessing the Spoofing Threat: Development of a Portable GPS Civilian Spoofer,"* Proceedings of the ION GNSS Meeting, (2008).

[5]. Mur-Artal, R., et al., *"ORB-SLAM: A Versatile and Accurate Monocular SLAM System,"* IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, (2015).

[6]. Lin, C., et al., *"A Survey of Large-Scale Image Localization Methods for UAVs,"* Journal of Computer Vision and Image Understanding, vol. 160, pp. 48-65, (2017).

[7]. Rublee, E., et al., *"ORB: An Efficient Alternative to SIFT or SURF,"* 2011 International Conference on Computer Vision, pp. 2564-2571, (2011).

[8]. Lowe, D. G., *"Distinctive Image Features from Scale-Invariant Keypoints,"* International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, (2004).

[9]. Ma, Y., et al., *"Image Registration Techniques: A Survey,"* IEEE Transactions on Medical Imaging, vol. 35, no. 7, pp. 1714-1735, (2016).

[10]. Johnson, J., Douze, M., and Je´gou, H., *"Billion-scale similarity search with GPUs,"* IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535-547, (2021).

[11]. Lowe, D. G., *"Object Recognition from Local Scale-Invariant Features,"* Proceedings of the International Conference on Computer Vision (ICCV), vol. 2, pp. 1150-1157, (1999).

[12]. Fischler, M. A., and Bolles, R. C., *"Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,"* Communications of the ACM, vol. 24, no. 6, pp. 381-395, (1981).

[13]. Kingma, D. P., and Welling, M., *"Auto-Encoding Variational Bayes,"* International Conference on Learning Representations (ICLR), (2014).

[14]. Douze, M., et al., *"The FAISS Library for Efficient Similarity Search,"* arXiv preprint arXiv:1702.08734, (2017).

[15]. Bay, H., Tuytelaars, T., and Van Gool, L., *"SURF: Speeded Up Robust Features,"* European Conference on Computer Vision (ECCV), (2006).

[16]. Bing, L. I., et al., *"UAV Image Matching Based on Improving the ORB Algorithm,"* Bulletin of Surveying and Mapping, (2024).

[17]. Mughal, M. H., Khokhar, M. J., and Shahzad, M., *"Assisting UAV Localization via Deep Contextual Image Matching,"* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 2445-2457, (2021).

[18]. Li, Q., et al., *"GeoFormer: An Effective Transformer-based Siamese Network for UAV Geo-localization,"* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, (2024).

[19]. Vaswani, A., et al., *"Attention is All You Need,"* Advances in Neural Information Processing Systems, (2017).

[20]. Bianchi, M., and Barfoot, T. D., *"UAV Localization Using Autoencoded Satellite Images,"* IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 1761-1768, (2021).
[21]. Di Piazza, T., et al., *"Leveraging Edge Detection and Neural Networks for Better UAV Localization,"* arXiv preprint arXiv:2404.06207, (2024).
[22]. Hou, X., et al., *"Deep Feature Consistent Variational Autoencoder,"* 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1133-1141, (2016).

TÓM TẮT

Phương pháp định vị hình ảnh hiệu quả cho UAV kết hợp Autoencoder và SIFT

Trong các môi trường không có hệ thống định vị vệ tinh toàn cầu (GNSS), việc đảm bảo định vị chính xác và đáng tin cậy cho máy bay không người lái (UAV) là một thách thức lớn. Bài báo này giới thiệu một phương pháp định vị dựa trên hình ảnh, kết hợp các kỹ thuật autoencoder và SIFT, gọi tắt là AE+SIFT. Phương pháp này nén các hình ảnh bản đồ độ phân giải cao thành các vector kích thước nhỏ, sau đó lưu trữ trong cơ sở dữ liệu để dễ dàng truy xuất. Trong quá trình định vị, hình ảnh UAV được mã hóa và so khớp với cơ sở dữ liệu, sau đó sử dụng SIFT và phép chiếu homography để định vị chính xác. Phương pháp AE+SIFT cải thiện độ chính xác định vị, đạt sai số tọa độ trung bình là 3,94 mét so với đường bay thực tế khi ảnh UAV cùng hướng với ảnh tham chiếu. Đặc biệt, khi hình ảnh UAV không cùng hướng với ảnh tham chiếu, phương pháp của chúng tôi vượt trội hơn so với phương pháp AE về độ chính xác định vị.

Từ khoá: GNSS; Autoencoder; SIFT; Định vị hình ảnh; UAV.