Image enhancement for UAV visual SLAM applications: analysis and evaluation.

TIAN, Y., YUE, H. and REN, J.

2024

Image Enhancement for UAV Visual SLAM Applications: Analysis and Evaluation

Yikun Tian[1], Hong Yue[1(✉)], and Jinchang Ren[2]

[1] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK
{yikun.tian,hong.yue}@strath.ac.uk
[2] National Subsea Centre, Robert Gordon University, Aberdeen AB21 0BH, UK
j.ren@rgu.ac.uk

Abstract. Although simultaneous localisation and mapping (SLAM) has been widely applied in a wide range of robotics and navigation applications, its applicability is severely affected by the quality of the acquired images, especially for those in unmanned aerial vehicles (UAV). In this paper, comprehensive analysis and evaluation of the methods for enhancement of the UAV images are focused, especially the models for denoising of the UAV images using spatial-domain analysis, transform domain analysis and deep learning. Experiments on publicly available datasets are conducted for performance evaluation, along with both qualitative and quantitative results. Surprisingly, deep learning-based approaches did not perform particularly well as these did in other computer vision tasks such as object detection and recognition. Useful discussions are suggested how to further explore this interesting topic.

Keywords: Unmanned Aerial Vehicle (UAV) · visual SLAM · image enhancement · denoising · dehazing

1 Introduction

1.1 A Subsection Sample

With the rapid development of the unmanned aerial vehicle (UAV) techniques, there is a growing trend to apply it to a wide range of applications, including but not limited to survey, surveillance and inspection, supporting various industrial, civil, agriculture, energy, transportation and military tasks. Within these tasks, autonomous visual navigation is a fundamental requirement to enable the automated pilot and survey, for which the simultaneous location and mapping (SLAM) has been widely applied for decades.

There are two major tasks in UAV based visual SLAM, which are constructing a map of the surrounding environment and accurately estimating the motion trajectory of the UAV itself. However, when the UAV system works outdoors, affected by factors such as weather and poor/inconsistent light, the quality of the aerial images on which its visual SLAM environment map construction relies will be seriously degraded. Actually, the

quality of the images acquired from UAV platforms may severely affect the accuracy and efficacy of SLAM based navigation tasks. As a result, it will not be possible to accurately estimate the UAV's motion attitude trajectory through aerial images, and it will not be possible to complete the construction of the visual SLAM environment map.

For accurately constructing the surrounding environment map, denoising and enhancement of the image collected by the UAV system becomes essential, before constructing the environment map of visual SLAM. To date, many different models and approaches have been proposed for denoising of aerial images. It is our aim to provide a useful survey and comprehensive evaluation of these models, which will provide a strong base for researchers working in the area to choose the best models accordingly.

2 Image Denoising Models

For denoising of aerial images, numerous researchers have proposed a number of image denoising methods, primarily categorized into traditional image denoising and deep learning-based image denoising methods. Traditional denoising methods can be further subdivided into spatial domain denoising methods and transform domain denoising methods, as detailed below.

2.1 Spatial-Domain Denoising Models

Spatial domain methods primarily utilize filters for denoising. They process the neighbourhood of each pixel in the image using a filter, iterating through the entire image. Spatial domain denoising methods can be classified based on the linearity of the filters into linear filtering methods and non-linear filtering methods [1].

In linear filtering methods, the most common one is the mean filter. For a pixel contaminated by noise, the mean filter calculates the average value of all the pixels in its neighbourhood and assigns this average value to the contaminated pixel. Non-linear filtering methods typically include the median filter and bilateral filter. The median filter initially sorts the pixels around a particular pixel, resulting in an ordered data sequence. Then, it assigns the median value from this sequence to the pixel, effectively removing low and high-frequency components in noisy images. Thus, it is commonly used for eliminating salt-and-pepper noise. However, it has the drawback of potentially causing image discontinuities.

The bilateral filter [2] considers both the grayscale similarity and spatial position relationships between pixels. It assigns higher weight values to pixels that are both close to the center pixel and have similar grayscale values, while giving lower weight values to pixels that are farther away or have dissimilar grayscale values. The advantage of the bilateral filter lies in its ability to preserve more edge information, but it requires further improvement in protecting image texture and detail information.

Local filters can effectively remove noise when the noise level is relatively low but are less effective at higher noise levels. To address this issue, the Non-Local Means (NLM) [3] denoising algorithm leverages the self-similarity and redundancy in the image's structure for denoising. Danbov et al. [4] introduced the Block Matching 3D (BM3D)

denoising algorithm, which involves finding a series of similar image blocks and grouping them to obtain multiple three-dimensional blocks. Filtering is then performed in three-dimensional space, followed by using a three-dimensional inverse transform to produce the denoised result. Compared to NLM, this algorithm achieves a higher peak signal-to-noise ratio (PSNR) but comes with higher complexity.

2.2 Transform-Domain Denoising Models

Transform domain methods exploit the distinctive characteristics of images and noise in the transform domain to perform denoising. In the early stages of development for transform domain denoising methods, Fourier transformation was used to remove noise from images. Fourier transformation converts data from the time domain to the frequency domain, where noise in the frequency domain often appears in high-frequency regions. Noise removal can be achieved through low pass filtering in the frequency domain. However, this process also eliminates the texture and detail information in the image.

In addition to Fourier transformation, wavelet transformation has also been employed for image denoising. Denoising methods utilizing wavelet transformation process noise removal based on the differences between image features and noise after undergoing wavelet transformation. The advantage of wavelet-based denoising is that it can simultaneously preserve both frequency and spatial information in the image. However, its drawback lies in its weaker directionality, as it can only extract limited directional information.

2.3 Deep Learning Based Denoising Models

In recent years, deep learning has gained the favor of many researchers due to its powerful feature capturing capabilities and flexible network architectures. Burger et al. employed a Multi-Layer Perceptron (MLP) [5] to learn the mapping from noisy images to clean images, achieving performance comparable to BM3D. Chen et al. [6] designed a trainable Nonlinear Reaction Diffusion (TNRD) denoising model. However, MLP and TNRD can only handle images with fixed noise levels and may not yield ideal results when applied to datasets with varying noise levels.

To enhance the model's ability to handle varying levels of noise, Zhang et al. [7] introduced the Denoising Convolutional Neural Network (DnCNN) model. This model not only addresses images with different noise levels but also utilizes residual learning and batch normalization techniques to expedite model training. Subsequently, Zhang et al. proposed the Fast and Flexible Denoising Network (FFDNet) model [8], which builds upon DnCNN by including noise levels as an additional input to the model. FFDNet is capable of handling spatially correlated noise and plays a crucial role in balancing noise reduction and image detail preservation.

However, these aforementioned models do not yield satisfactory results for real image denoising. To tackle this issue, Guo et al. [9] introduced the Convolutional Blind Denoising Network (CBDNet) and constructed a new noise model to simulate real noise. CBDNet consists of a denoising sub-network and a noise level estimation sub-network, enhancing the network's performance and generalization capacity by introducing an asymmetric loss function. Nevertheless, CBDNet's network structure is complex and

comes with a significant computational cost, making it less suitable for practical applications. Therefore, Anwar and Barnes [10] proposed the Real Image Denoising Network (RIDNet) to address real-world denoising scenarios. RIDNet adopts a modular structure for the denoising network and introduces a channel attention mechanism for adaptive channel weight adjustment. The introduction of CBDNet and RIDNet has driven the development of image denoising research in real-world settings.

3 Datasets and Evaluation Criteria

3.1 Datasets Description

For image denoising, the CBSD68 [11] and the SIDD [12] datasets were used. The CBSD68 dataset consists of 68 colour images of varying sizes. The SIDD dataset includes approximately 30,000 noisy images captured under different lighting conditions using five representative smartphones, along with corresponding "noise-free" ground truth images. In addition, an own dataset including 2035 virtual simulation scene images was also used for testing the effect of denoising.

For image dehazing assessment, the SOTS-outdoor [13] public dataset is used. The SOTS dataset is a synthetic dataset consisting of 1000 test images, divided into indoor and outdoor categories, each containing 500 images.

3.2 Evaluation Metrics

3.2.1 Peak Signal-to-Noise Ratio (PSNR)

Given a hazy image I and a haze-free image K both of size M*N, the Mean Squared Error (MSE) is defined as:

$$MSE = \frac{1}{MN} \sum_{j=0}^{M-1} \sum_{j=0}^{N-1} \left[I(i,j) - K(i,j) \right]^2 \tag{1}$$

where $I(i,j)$ and $K(i,j)$ represent the grayscale values of pixels at location (i,j) in the hazy and haze-free images, respectively. The Peak Signal-to-Noise Ratio (PSNR) is then defined as

$$PSNR = 10log_{10} \left(\frac{(maxI)^2}{MSE} \right) \tag{2}$$

where MAX represents the maximum possible pixel value in the image. This formula is commonly used for grayscale images. For colour images with three channels (RGB), the MSE is calculated separately for each channel, and the resulting MSE values are used to compute the PSNR for each channel. The final PSNR for the color image is obtained by taking the average of the PSNR values across all channels.

3.2.2 Structural Similarity (SSIM)

SSIM is used to measure the structural similarity between two images, which compares the structure, luminance, and contrast of two images. For the two given images x and y, the structural similarity between the two images can be computed as follows [14].

$$SSIM\,(x, y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{3}$$

where μ_x represents the average value (mean) of x, μ_y is the average value (mean) of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y. The constants $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are used to maintain stability and are typically set to small values like 0.01 and 0.03, respectively. The range of the Structural Similarity (SSIM) metric is from -1 to 1, with larger values indicating less distortion. When two images are identical, the SSIM value is 1.

4 Results and Analysis

4.1 Compared Methods

Based on the recommended good performance, the following models are selected for evaluation in our experiments.

- **BM3D** [4]**,** as a novel image denoising method, BM3D is based on an enhanced sparse representation in the transform domain. The enhanced sparsity is achieved by grouping similar 2-D fragments into 3-D data arrays namely "groups", followed by collaborative filtering being applied to these 3-D groups. The collaborative filtering has helped to reveal the finest details shared by grouped fragments whilst preserving the essential unique features of each individual fragment.
- **Weighted Nuclear Norm Minimization** (WNNM) [15]: The image is modeled as Y = X - N, where Y is also composed of samples with noise, forming a sample matrix. X and N are the corresponding noise-free sample matrices and noise, respectively. The given constraint is that X is a low-rank matrix. Because the matrix composed of similar samples exhibits low-rank characteristics, while the noise does not have low-rank characteristics, image denoising can be achieved through low-rank clustering.
- **Variational Denoising Network** (VDN) [16]: This model is capable of simultaneous image denoising and noise estimation. In typical work, Gaussian white noise is assumed to be present in the image, but this model is not limited to that. The proposed generative model exhibits strong generalization capabilities and performs well even for noise not encountered in the test set. The model provides an explanation for the overfitting phenomenon often observed in deep learning methods trained using MSE loss. This issue is attributed to overfitting the prior of the underlying clean image while neglecting variations in noise. This model explicitly models the generation of noise, thereby avoiding this drawback of deep learning methods.

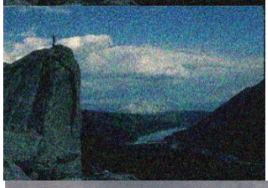
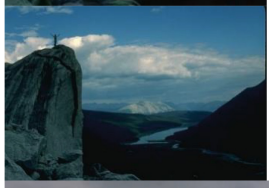
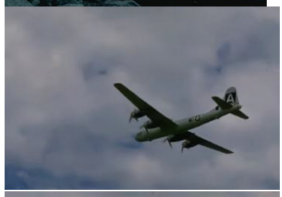
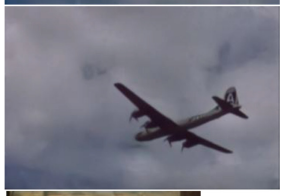
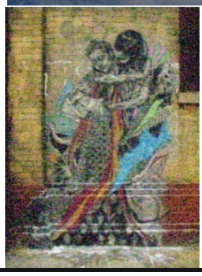
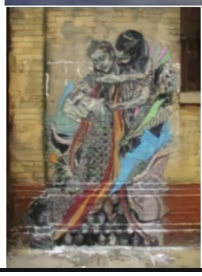
- **FFDNet** [8]: The adjustable noise level mapping, denoted as M, is used as input to provide flexibility to the denoising model regarding noise levels. An invertible downsampling operator is introduced to reshape the input image of size $W \times H \times C$ into four subsampled sub-images of size $4W/2 \times H/2 \times 4C$, where C represents the number of channels. To ensure that noise level mapping robustly controls the trade-off between denoising and detail preservation without introducing visual artifacts, an orthogonal initialization method is applied to the convolution filters.
- **NAFNet** [17]: Taking inspiration from the Transformer architecture, the use of Layer Normalization (LN) is incorporated to facilitate smoother training. NAFNet also introduces LN operations, leading to significant performance gains on image denoising and deblurring datasets. In the Baseline approach, ReLU is jointly replaced with GELU and CA. GELU helps maintain denoising performance while significantly enhancing deblurring performance. Two new attention module compositions are proposed, namely CA (Channel Attention) and SCA (Spatial-Channel Attention).
- **CycleISP** [18]: The images captured by the camera initially exist as RAW-RGB images, where each pixel contains only one of the three-color channels: R, G, or B. These RAW images are then processed through the camera's ISP (Image Signal Processing), which includes operations like noise reduction, white balance adjustment, gamma transformation, tone mapping, and more, resulting in sRGB images (standard-RGB with three channels). Two neural networks have been employed to simulate this process in both forward and reverse directions. In other words, these networks can transform sRGB images into RAW images and vice versa.

4.2 Results and Analysis

Table 1 below presents a comparison of image denoising experiment results using different algorithms on publicly available datasets. The BM3D [4], WNNM [15], VDN [16], FFDNet [8], and NAFNet [17] were tested on the CBSD68 dataset, while the CycleISP [18] method was applied to the SIDD [12] dataset. The PSNR and SSIM values in the table above represent the peak performance achieved by the respective algorithms on this dataset. From the results, it can be observed that the BM3D and NAFNet algorithms perform well in denoising based on publicly available datasets.

As seen, BM3D apparently outperforms WNNM, thanks for considering the local self-similarity, which has greatly improved the performance of denoising, including those using deep learning models. The relatively poor performance from deep learning can be due to two reasons, i.e. insufficient training and inconsistency between the training and testing samples. The latter may be caused by the random characteristics of the noise within the image, which has potentially affected the learning-based approaches. Nevertheless, in the best deep learning model, NAFNet, the transformer architecture has somehow mitigated such limitations, which can be further explored.

Table 1. Comparison of Image Denoising Methods Using Public Datasets.

Methods	Results	Original Image	Resulted Image
BM3D	PSNR: 41.63dB SSIM: 0.9936		
WNNM	PSNR: 39.38dB SSIM: 0.9750		
VDN	PSNR: 30.83dB SSIM: 0.8533		
FFDNet	PSNR: 28.98dB SSIM: 0.7969		
NAFNet	PSNR: 40.30dB SSIM: 0.9621		
CycleISP	PSNR: 34.70dB SSIM: 0.9822		

5 Conclusions

In this paper, a survey of the denoising models for UAV images in SLAM implementation is focused, followed by a comprehensive evaluation. Six models are selected for both qualitative and quantitative assessment, including conventional approaches in the spatial

domain and transform domain as well as deep learning models. By benchmarking on the publicly available datasets, it is found that the BM3D model outperforms all others, even the deep learning approaches, owing mainly to the local-similarity being used in modelling.

This one hand shows the great potential of conventional vision - based perception models in image denoising. On the other hand, it indicates the potential limitations of the deep learning models in this context, due mainly to the ill-posed problem in training the models. Furthermore, the great potential of the NAFNet has suggested that the transformer architecture can help to mitigate the limitations here and improve the modeling thus is worth further investigation.

As the degradation process of UAV images can be much more complicated [19], this paper only covers a small portion, where many other useful topics have not been covered, such as image dehazing, deblurring and normalisation of the lighting effects et al. These will be our future work, along with the integration of other challenging models.

References

1. Cai, R.: Research progress in image denoising algorithms based on deep learning. J. Phys. **1345**, 042055 (2019)
2. Banterle, F., Corsini, M., Cignoni, P., Scopigno, R.: A low-memory, straightforward and fast bilateral filter through subsampling in spatial domain. Comput. Graph. Forum **31**(1), 19–32 (2011)
3. Buades, A.: A non-local algorithm for image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 60–65 (2005)
4. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
5. Burger, H.C., Schuler, C.J., Harmeling, S.: Image denoising with multi-layer perceptrons, part 1: comparison with existing algorithms and with bounds. Comput. Sci. 8–30 (2012)
6. Chen, W., Huang, Z., Tsai, C., et al.: Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: toward a unified model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17653–17662 (2022)
7. Zhang, K., Zuo, W., Chen, Y., et al.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
8. Zhang, K., Zuo, W., Zhang, L.: FFDNet: toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)
9. Guo, S., Yan, Z., et al.: Toward convolutional blind denoising of real photographs. In: Proceedings of the CVPR, pp. 1712–1722 (2019)
10. Anwar, S., Barnes, N.: Real image denoising with feature attention. In: Proceedings of the ICCV, IEEE, pp. 3155–3164 (2019)
11. Martin, D., Fowlkes, C., et al.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the 8th IEEE International Conference on Computer Vision, pp. 416–423 (2001)
12. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1692–1700 (2018)

13. Hu, X., Fu, C., Zhu, L., et al.: Direction-aware spatial context features for shadow detection and removal. IEEE Trans. Pattern Anal. Mach. Intell. **42**(11), 2795–2808 (2019)
14. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
15. Gu, S., Zhang, L., et al.: Weighted nuclear norm minimization with application to image denoising. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2862–2869 (2014)
16. Yue, Z., Yong, J., Zhao, Q., Meng, D., Zhang, L.: Variational denoising network: toward blind noise modeling and removal. Adv. Neural Inf. Process. Syst. **32** (2019)
17. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. LNCS, vol. 13667, pp. 17–33. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_2
18. Zamir, S.W., et al. :CycleISP: real image restoration via improved data synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2693–2702 (2020)
19. Fu, H., et al.: Three-dimensional singular spectrum analysis for precise land cover classification from UAV-borne hyperspectral benchmark datasets. ISPRS J. Photogram. Remote Sens. **203**, 115–134 (2023)