Innovative Fusion of Transformer Models with SIFT for Superior Panorama Stitching

Zheng Xiang

School of Biotechnology, Jiangnan University, Wuxi, 214122, China
100366@yzpc.edu.cn

Abstract. In the field of image stitching, generating multiple panoramas from a large set of images is a challenging task. Traditional methods often require complex pairwise comparisons, leading to time-consuming operations that may affect accuracy and efficiency. To address this issue, this paper presents an innovative method aimed at improving computational efficiency for generating multiple panoramas in multi-class grouping. By introducing vision transformer models and cosine similarity metric, our approach enables rapid evaluation of relationships between image pairs in the initial phase, thus reducing dataset size and minimizing time-consuming feature matching operations. By initially utilizing vision transformer models to extract features from each image, we implement a cosine similarity metric to rapidly assess preliminary relationships between image pairs. This preliminary phase allows for the reduction of the dataset subjected to the more computationally intensive, Fast Library for Approximate Nearest Neighbors-based (FLANN) feature matching. Experimental results demonstrate that our method achieves a 93.34% reduction in computational time compared to traditional methods, with only an 8.12% decrease in clustering accuracy. This improvement is attributed to the effective utilization of preliminary relationship assessments to optimize the feature matching process and achieve a more efficient generation of multiple panoramas in multi-class grouping.

Keywords: Panorama Generation, Vision Transformers, Feature Matching, Image Clustering

1 Introduction

In the realm of digital imaging, panorama generation and autostitching stand as critical components, particularly in applications spanning from virtual reality to geographical information systems. Traditional image stitching techniques aim to seamlessly merge multiple overlapping images into a single composite panorama without noticeable distortions or seams. Key to these methods has been the development of robust feature detection algorithms like Scale-Invariant Feature Points (SIFT), introduced by Lowe [1], which remain foundational in identifying and matching features across different images reliably.

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_78

Recent advancements in digital imaging have explored the incorporation of machine learning techniques, particularly convolutional neural networks (CNNs), which have been employed to automate and enhance the stitching process [2]. However, the introduction of Vision Transformers (ViTs) has revolutionized various domains of computer vision, including image classification. Transformers, initially designed for natural language processing tasks, were adapted for image recognition tasks by Dosovitskiy et al. [3], demonstrating remarkable capabilities in handling intricate visual data through self-attention mechanisms. This adaptation has opened new pathways for their application in image processing tasks beyond typical classification challenges.

Building upon these advancements, my work integrates the deep learning capabilities of Vision Transformers with traditional computer vision techniques such as SIFT and Fast Library for Approximate Nearest Neighbors (FLANN) based matching, introduced by Muja and Lowe [4]. This integration aims to address the inherent challenges of feature mismatch and image misalignment in traditional stitching methods. By combining the contextual understanding capabilities of transformers with the precise local feature matching of SIFT, this approach enhances both the accuracy and efficiency of panorama generation. Studies like those by He et al. [2] and Bai et al. [5] illustrate ongoing efforts to refine image stitching and related tasks, underscoring the need for innovative methodologies that can handle diverse and complex imaging scenarios effectively [6].

Moreover, the potential of transformers in computer vision has been explored in various contexts, such as in the work by Ayana and Choe [7] for medical image classification and by Pucci et al. [8] for fine-grained image categorization. These studies highlight the versatility and robustness of transformer models in extracting and processing high-level features from images, supporting their suitability for tasks that require a nuanced understanding of visual content, such as panorama stitching.

Thus, this research not only demonstrates the potential of transformers in computer vision tasks beyond their conventional uses but also proposes a methodological framework for their application in automating panorama generation with high precision. By leveraging the advanced capabilities of pre-trained transformer models alongside proven computer vision algorithms, this research aims to surpass the limitations of conventional methods, enhancing the quality and functionality of panoramic images.

2 Method

In the field of image processing, creating high-quality panoramas from multiple overlapping images remains a challenging task. Traditional methods often struggle with efficient and accurate feature extraction, particularly in complex environments or with large datasets. These limitations primarily arise from their reliance on hand-crafted features, which may not adequately capture the global context of images, leading to suboptimal stitching results.

This paper seeks to address these shortcomings by integrating ViTs at the initial phase of our approach. ViTs are particularly adept at extracting rich, global feature

representations from images, making them suitable for identifying potential overlaps in panorama stitching (Figure 1). We hypothesize that adjacent image pairs with substantial overlaps will exhibit high similarity in their vector features, which can be effectively captured by ViTs. This capability allows us to overcome the limitations of traditional methods that often fail to recognize complex patterns without extensive preprocessing.

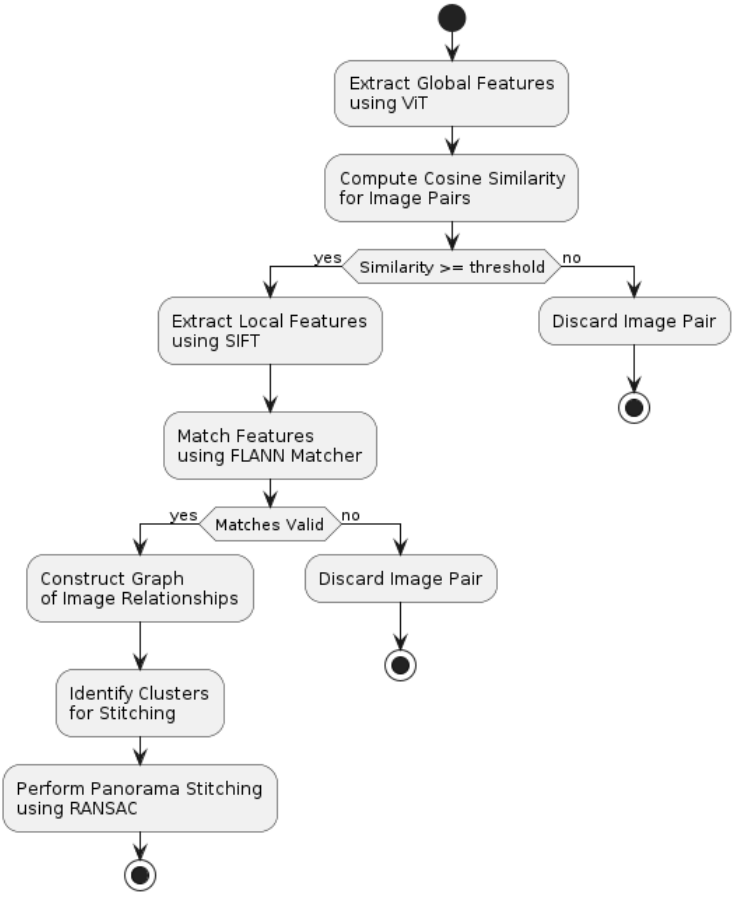

Fig. 1. Processing logic diagram

To complement the global perspective provided by ViTs, SIFT and FLANN are employed in subsequent stages for precise local feature detection and matching. This combination ensures that the initial assessments made by the ViTs are verified and refined, leading to accurate alignment of images in the final panorama.

Furthermore, to validate the effectiveness of our approach, we utilize the adjusted Rand score. This robust statistical tool serves as a measure of the quality of our image groupings, offering a normalized assessment that accounts for the possibility of random chance in the formation of these groups. By adopting this comprehensive approach, our

research aims to significantly improve the accuracy and efficiency of panorama stitching, providing a meaningful advancement over existing methods.

2.1 Tools

Vision Transformers ViTs apply the principles of self-attention, originally used in natural language processing, to image analysis tasks. A ViT divides an image into patches, which are then flattened and processed through a series of transformer blocks. Each block consists of multi-head self-attention and feed-forward neural networks. The mathematical formulation of the self-attention mechanism can be represented as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})\mathrm{V} \tag{1}$$

where Q, K, V are the queries, keys, and values derived from the input patches, and dk is the scaling factor based on the dimension of the keys. This approach allows ViTs to capture global dependencies within an image.

SIFT SIFT identifies and describes local features in images. The process involves four major steps: scale-space extrema detection, keypoint localization, orientation assignment, and keypoint descriptor. The keypoint descriptor provides a unique and invariant fingerprint of the feature that is robust to changes in scale, orientation, and illumination.

FLANN FLANN is used for efficient search of nearest neighbors in large datasets. It works by building randomized kd-trees or hierarchical k-means trees to partition the data space and then traversing these trees to find the best matches. This method is particularly useful in reducing the computational cost of matching features in large-scale image datasets [4].

Adjusted Rand Score The adjusted Rand score, available through sklearn.metrics, is a measure of the similarity between two data clusterings, adjusted for chance. It is defined as:

$$Adjusted\ Rand\ Score = \frac{Index - Expected\ Index}{Max\ Index - Expected\ Index} \tag{2}$$

where the index is the Rand index, and the expected index is the expected value of the Rand index given random cluster assignments [9,10].

2.2 Implementation

Feature Extraction and Preliminary Assessment

- ViT Model: The pre-trained google/vit-base-patch16-224 Vision Transformer model is utilized for the initial extraction of global features from each image. These features are expected to reflect the extent of overlap between adjacent images, based on the hypothesis that pairs with sufficient overlap should share similar vector features.
- Similarity Measurement: Using the features extracted by the ViT model, we compute cosine similarity for each image pair. This step serves as a preliminary filter to determine potential candidates for further processing, based on set similarity thresholds (-1, 0.35, 0.4, 0.45, 0.5). The threshold of 0.35 aligns with the hypothesis that adjacent images should overlap by more than one-third of their size for effective stitching.

Detailed Feature Matching and Verification

- SIFT: If the similarity score between an image pair meets or exceeds the threshold, we proceed to extract detailed local features using SIFT. This method identifies stable features across varying scales and conditions, crucial for precise alignment.
- FLANN-Based Matching: After SIFT feature extraction, the FLANN SIFT: If the similarity score between an image pair meets or exceeds the threshold, we proceed to extract detailed local features using SIFT. This method identifies stable features across varying scales and conditions, crucial for precise alignment.matcher is employed to find and validate matches between the descriptors of the image pairs that passed the initial similarity filter. This stage uses Lowe's ratio test with a distance threshold of 0.6, ensuring that only significant matches are considered for stitching. This selective approach aims to reduce computational overhead by limiting FLANN matching to only the most promising image pairs.

Graph Construction and Panorama Formation

- Graphical Representation: A graph is constructed where nodes represent individual images, and edges—established based on the initial similarity assessment and validated by FLANN matches—connect images that are determined to be sufficiently similar. This graph aids in visualizing the sequence and relationships among images, facilitating effective clustering.
- Panorama Stitching: For each cluster identified in the graph, panorama stitching is performed using RANSAC to compute homography matrices, ensuring robust alignment of images. This stage focuses on achieving precise image alignment, stitching images together based on the computed transformations without incorporating advanced blending techniques.

3 Analysis and Results

3.1 Data and Results

The dataset consists of 140 photographs taken by the author, depicting diverse urban and natural environments such as street views, buildings, and plants. These images form the basis for testing the enhanced panorama generation model (Table 1).

Table 1. Impaction to images grouping of different similarity thresholds on different datasets

Dataset Size	Similarity Threshold	Adjusted Rand Score	Elapsed Time (seconds)	FLANN Compare Counts
34 Pictures	-1	91.13%	45.58	561
	0.35	91.13%	23.71	206
	0.4	91.13%	21.29	174
	0.45	91.13%	17.98	133
	0.5	83.17%	15.41	103
46 Pictures	-1	91.51%	123.14	1035
	0.35	94.49%	42.26	267
	0.4	94.49%	38.29	229
	0.45	94.49%	34.46	186
	0.5	89.66%	31.95	154
56 Pictures	-1	95.71%	179.79	1540
	0.35	95.71%	56.48	350
	0.4	95.71%	49.29	291
	0.45	95.71%	42.91	221
	0.5	91.98%	38.08	192
63 Pictures	-1	96.16%	211.67	1953
	0.35	96.16%	64.12	474
	0.4	96.16%	56.62	392
	0.45	96.16%	47.05	306
	0.5	92.83%	42.50	246
130 Pictures	-1	97.50%	1081.29	8385
	0.35	92.19%	180.16	1123
	0.4	92.19%	131.60	806
	0.45	90.74%	105.30	568
	0.5	82.18%	86.64	419
140 Pictures	-1	87.34%	1184.15	9730
	0.35	81.69%	188.27	1248
	0.4	81.69%	152.35	907
	0.45	80.25%	109.08	648
	0.5	71.48%	89.12	480

3.2 Analysis

The analysis of different similarity thresholds demonstrates a trade-off between computational efficiency and grouping accuracy, evaluated using the adjusted Rand score. The study investigates how varying the similarity threshold impacts the time cost and number of FLANN feature comparisons required across different dataset sizes (Table 1).

The results indicate that increasing the similarity threshold generally decreases the time cost and the number of comparisons, confirming our hypothesis that higher thresholds can reduce the computational burden. However, this comes at the expense of decreased grouping accuracy, especially at higher thresholds (Figure 2, figure 3, figure 4, and figure 5).

For smaller datasets (34 and 63 pictures), increasing the similarity threshold generally maintains a high Adjusted Rand Score, implying good clustering quality.

For larger datasets (130 and 140 pictures), a noticeable decline in clustering quality (Adjusted Rand Score) is observed as the threshold increases. This may indicate that a higher threshold could be too restrictive, leading to missed valid overlaps in larger datasets.

Notably, for the 140-picture dataset, setting the threshold at -1 results in a significant increase in both time cost and FLANN comparisons (1184.15 seconds and 9730 counts) compared to a threshold of 0.5 (89.12 seconds and 480 counts). This is indicative of the substantial computational savings achieved by applying a higher similarity threshold.

Higher thresholds generally result in quicker computations but at the potential cost of accuracy in clustering, as seen by a drop in the Adjusted Rand Score, especially noticeable in the 140-picture dataset.

Choosing 0.45 as the similarity threshold, our method achieves a 93.34% reduction in computational time compared to traditional methods, with only an 8.12% decrease in clustering accuracy.

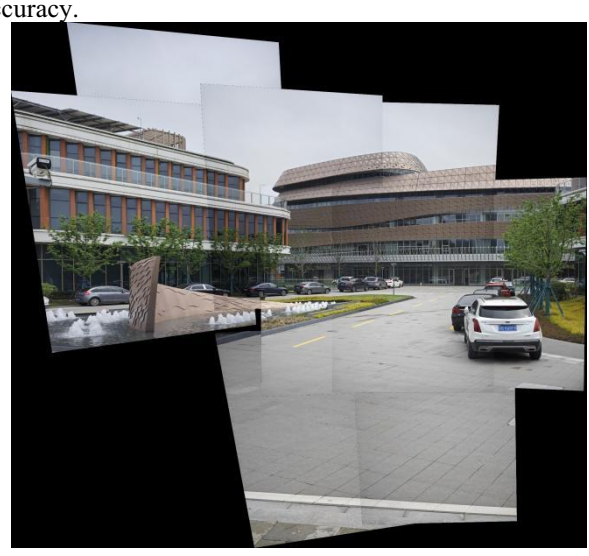

Fig. 2. Example of generated panorama

Fig. 3. Example of generated panorama

Fig. **4**. Example of generated panorama

Fig. 5. Example of generated panorama

4 Discussion

4.1 Implications of Findings

The results of this study illustrate a clear trade-off between computational efficiency and the accuracy of image groupings in panorama stitching. Increasing the similarity threshold significantly reduces both the computational time and the number of comparisons required, which aligns with the initial hypothesis that higher thresholds can streamline the feature matching process. This finding is particularly relevant in contexts where processing time and resource utilization are critical factors, such as in real-time panorama stitching applications or when processing large datasets on limited hardware.

However, the study also highlights the potential drawbacks of setting the threshold too high. While a threshold of 0.45 generally maintains an acceptable balance between efficiency and accuracy in smaller datasets, it leads to a noticeable degradation in performance as the dataset size increases, as seen with the 140 pictures dataset. This degradation is manifested in more granulated groups, which may not be desirable in applications requiring high fidelity in image stitching.

4.2 Considerations for Practical Application

For practical applications, the choice of threshold should consider the specific requirements of the task at hand. For instance, applications that prioritize speed over precision, such as quick previews in photo stitching software or applications in dynamic environments (e.g., drone imaging), might opt for higher thresholds. Conversely, applications where precision is paramount, such as in detailed geographic mappings or cultural heritage preservations, might require lower thresholds to ensure finer detail and accuracy.

4.3 Limitations

One limitation of the study is the focus on a specific set of thresholds and image datasets. Future studies could expand the range of thresholds tested and include a broader variety of image types and conditions to determine how these factors might influence the optimal threshold setting. Additionally, the study does not account for potential biases in the dataset selection, which might affect the generalizability of the results.

4.4 Future Research

Further research could explore the integration of adaptive thresholding mechanisms that adjust based on the content and characteristics of the image set, potentially enhancing both efficiency and accuracy dynamically. Moreover, investigating the combination of vision transformers with other types of feature extraction and matching algorithms could provide insights into more robust methods for panorama stitching. Another promising avenue would be the application of machine learning techniques to predict the optimal similarity threshold based on the initial assessment of image features, thereby automating and optimizing the stitching process further.

5 Conclusion

Driven by the need for efficient and accurate panorama stitching in the realm of computer vision, this research embarked on a journey to optimize the process through the innovative application of ViTs, combined with traditional methods like SIFT and FLANN. The work rigorously tested various similarity thresholds to understand their impact on computational load and clustering quality.

Our experiments have conclusively demonstrated that an increased similarity threshold correlates with reduced time costs and FLANN comparison counts, significantly enhancing computational efficiency across all dataset sizes. However, it's apparent that excessively high thresholds lead to diminished clustering accuracy, particularly for larger datasets. Notably, a threshold of 0.45 strikes an ideal balance for smaller datasets, markedly decreasing computational demands while preserving the integrity of image groupings. Larger datasets, in contrast, necessitate a more cautious approach to threshold setting to ensure clustering fidelity.

The study’s insights pave the way for future explorations into adaptive thresholding and the merging of vision transformers with varied feature detection methods for panorama stitching. Upcoming research should broaden the scope to include diverse image datasets and thresholds, and harness machine learning to refine and automate the stitching process.

References

1. D.G. Lowe, 2004. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 60(2), 91–110
2. X. He, L. He, et al., 2021. Image Stitching via Convolutional Neural Network. In: 7th International Conference on Computing and Data Engineering, 1-10. Springer, Heidelberg
3. A. Dosovitskiy, et al., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929
4. M. Muja, D.G. Lowe, 2009. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. Int. J. Comput. Vis. 85(2), 123–137
5. J. Bai, Z. He, et al., 2023. Local-to-Global Panorama Inpainting for Locale-Aware Indoor Lighting Prediction. IEEE Transactions on Visualization and Computer Graphics29(4), 1375-1384
6. Q. Wang, W. Li, et al., 2024. 360DVD: Controllable Panorama Video Generation with 360-Degree Video Diffusion Model. arXiv:2401.06578
7. G. Ayana, S. Choe, 2024. Vision Transformers-Based Transfer Learning for Breast Mass Classification From Multiple Diagnostic Modalities. J. Electr. Eng. Technol. 19(1), 234-245
8. R. Pucci, V.J. Kalkman, D. Stowell, 2024. Performance of computer vision algorithms for fine-grained classification using crowdsourced insect images. arXiv:2404.03474
9. L. Hubert, P. Arabie, 1985. Comparing Partitions. J. Classif. 2(1), 193–218
10. L. Wu, X. Lin, Z. Chen, et al., 2020. Surface crack detection based on image stitching and transfer learning with pretrained convolutional neural network. Struct. Control Health Monit., DOI:10.1002/stc.2766