3D Scene Reconstruction from Assorted 2D Images

Kat He
Columbia University
New York, NY 10027
kh3030@columbia.edu

Sae Na Na
Columbia University
New York, NY 10027
sn2943@columbia.edu

Abstract

This paper proposes and explores multiple modifications to the standard Structure from Motion (SfM) pipeline used for constructing 3D models from 2D images. Our goal is to seek methods that enable us to create precise 3D maps from real-life, diverse image sets that include six different challenges: phototourism, temporal changes, mixed aerial-ground captures, repeated structures, natural environments, and transparencies. Our approach integrates traditional image matching with modern machine learning techniques, image pre-processing techniques, and some novel ideas on enhancing the accuracy and robustness of 3D reconstructions. We conclude that image rotation correction lead to significant improvement multiple image resolutions showed improvement in runtime but maintained the same score, and our other proposals did not lead to quantifiable success, but further studies are proposed to dig into these in the future.

1. Introduction

In the field of computer vision, the creation of accurate three-dimension (3D) spatial models from sets of two-dimensional (2D) images is a challenge that continues to bring about significant interest, with Structure from Motion (SfM) being the notable solution in this domain. Traditionally, SfM relies on homogeneously high-quality data often obtained under controlled conditions [10]. However, real-world images are very often not taken in such ideal conditions. In order for SfM to develop more toward real-world applicability, there needs to be methods that are flexible and robust against a wide array of imaging conditions.

Specifically, this report explores solutions for the Image Matching Challenge of CVPR 2024, using the provided dataset [3]. We are tasked with developing methods capable of generating accurate spatial representations from a sizeable amount of images of any given location taken under diverse conditions, at different angles, from drones, in dense forests, and during different lighting conditions such as day and night. There are six main focus categories, each reflecting critical real-world challenges: (1) Phototourism and historical preservation, (2) Night vs day and temporal changes, (3) Aerial and mixed aerial-ground, (4) Repeated structures, (5) Natural environments, and (6) Transparencies and reflections.

We aim to address these difficulties by synthesizing traditional image matching techniques with cutting-edge machine learning algorithms. To be able to leverage this assorted dataset to create a 3D reconstruction of a scene, we attempt to explore and improve upon the state-of-the-art Structure-from-Motion (SfM) [12] [13] methods from aiding navigation and mapping for autonomous devices, medical diagnosis of wound depth, and creating 3D models of real-world sites for archaeological study or for basis in product design. This approach intends to work with diverse image sets and also fosters a deeper understanding of the complexities involved in practical SfM applications. We seek to investigate the boundaries of what is achievable with SfM, enhancing its applicability in many fields. This paper will reference results from our code referenced here [7].

2. Related Work

The development and application of Structure from Motion (SfM) techniques represent a cornerstone in the evolution of computer vision, particularly in the context of 3D reconstruction from image collections. At the center of this process lies the problem of image registration, which involves identifying correspondences between local features—distinct keypoints within images that capture the same physical points of a scene. Once these correspondences are found, the 3D coordinates of the points can be triangulated to reconstruct the scene geometry.

Historically, techniques such as Scale-Invariant Feature Transform (SIFT) [9] developed in 1999, and Random Sample Consensus (RANSAC) [4] introduced in 1981, have been foundational in the robust matching of these local features under varying conditions. These traditional methods have shown resilience and often outperform newer ma-

chine learning approaches, especially in challenging scenarios where the volume and complexity of the data significantly increase. Despite the success of these traditional approaches, they misrepresent the accuracy of these methods in practical applications.

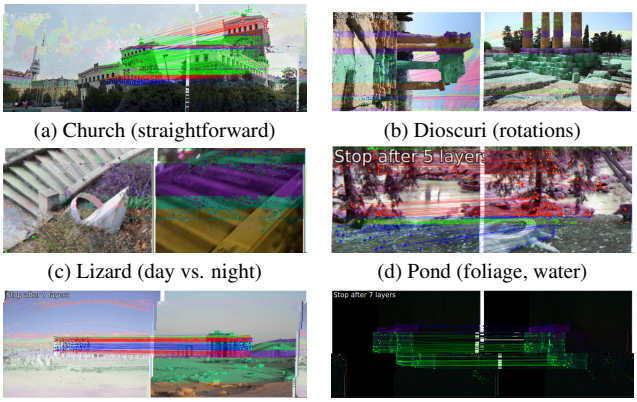
The current challenge is the third iteration of tackling the problem of reconstructing a 3D scene from a set of 2D images, but with the added dimension of integrating diverse environmental conditions and varied imaging scenarios into the task. This new dimension will further reduce the gap between the idealized problem statement to realistic applications of image reconstruction.

3. Methodology

3.1. Dataset

The training dataset consists of 2148 images of the 7 different scenes: church, dioscuri, lizard, temple, pond, transparent cup, and transparent cylinder, each representing the six main focus categories previously mentioned. The images in their respective scenes are non-homogeneous and are used together to create a 3D reconstruction. The test set consists of 1000 images of various scenes, representing the same six main focus areas. It is important to note that there is no overlap between the training set and test set. The training set is simply a representation of images used to illustrate the diverse of the diverse set of focus categories to challenge the robustness of our methods. Additionally, the training data has a sequential capture ordering and significant image-to-image content overlap while the test set has limited image-to-image overlap and the image ordering is randomized. The ground truth for the training data is a 3x3 rotation matrix and 3x1 translation vector representing the camera location with respect to ground truth, and the reference 3D reconstruction file.

Upon performing exploratory data analysis, we found that each dataset represented a variety of different challenges as shown in Figure 1. To identify which areas of this pipeline contribute to performance improvement, we visualized the scenes in the dataset by applying a simple keypoint detection and matching method. Figure 1b shows that the dioscuri dataset contains rotated images that confuse the keypoint matcher, as the model seems to expect vertical features to remain vertical in the image being matched. Figure 1c shows the difficulty in matching 2 images of the same staircase taken in the day and night, with the night image including some blur, possibly due to slow shutter speed. For this pair, only 3 keypoints were matched. This lead us to consider pre-processing the images to improve the blur, grain, and light balance in the dataset for better results. In contrast, Figure 1e shows 2 images with different lighting where almost 100 feature keypoints were able to be matched. This may be because despite the difference in lighting, both images still have high image quality (low blur). Through these observations, we developed our methodology to tackle each challenge.



(a) Church (straightforward) (b) Dioscuri (rotations)

(c) Lizard (day vs. night) (d) Pond (foliage, water)

(e) Temple scene (temporal shifts) (f) Cup (transparency)

Figure 1. Challenges presented by the dataset

3.2. Performance Metric

Submissions are evaluated on the mean Average Accuracy (mAA) of the registered camera centers $\mathbf{C} = -\mathbf{R}^{\mathbf{T}}\mathbf{T}$. The mean Average Accuracy (mAA) for a scene is the percentage of cameras, excluding the initial triplets, that are successfully registered using the optimized transformation T. The mAA for a scene is then averaged over several thresholds t_i ranging from approximately 1 cm to 1 m, which likely accommodate different precision requirements depending on the scene complexity with the final score obtained by averaging the mAA across all scenes in the dataset. This provides an overall measure of how well the algorithm performed in registering the camera centers across the diverse scenarios present in the dataset.

Each camera in the dataset is represented by its rotation matrix $\mathbf{R}$ and translation vector $\mathbf{T}$. The camera is considered "registered" if its calculated center $\mathbf{C}$ is within a certain threshold of the ground-truth center $\mathbf{C_g}$: mathematically, the norm of the difference between the ground-truth center and the transformed center is less than the threshold t: $||\mathbf{C_g} - \mathbf{T}(\mathbf{C})|| < \mathbf{t}$.

3.3. SfM Pipeline

Expanding on previous works that involve mainly keypoint detection and matching, we ideally wanted to focus on techniques that would improve the dataset to simplify the problem for the models when performing keypoint detection and matching. We also investigated models that would best work in conjunction with the pre-processing pipeline proposed in Figure 2.

In the initial pre-processing section of the pipeline, since the test dataset had images that were not in sequential cap-

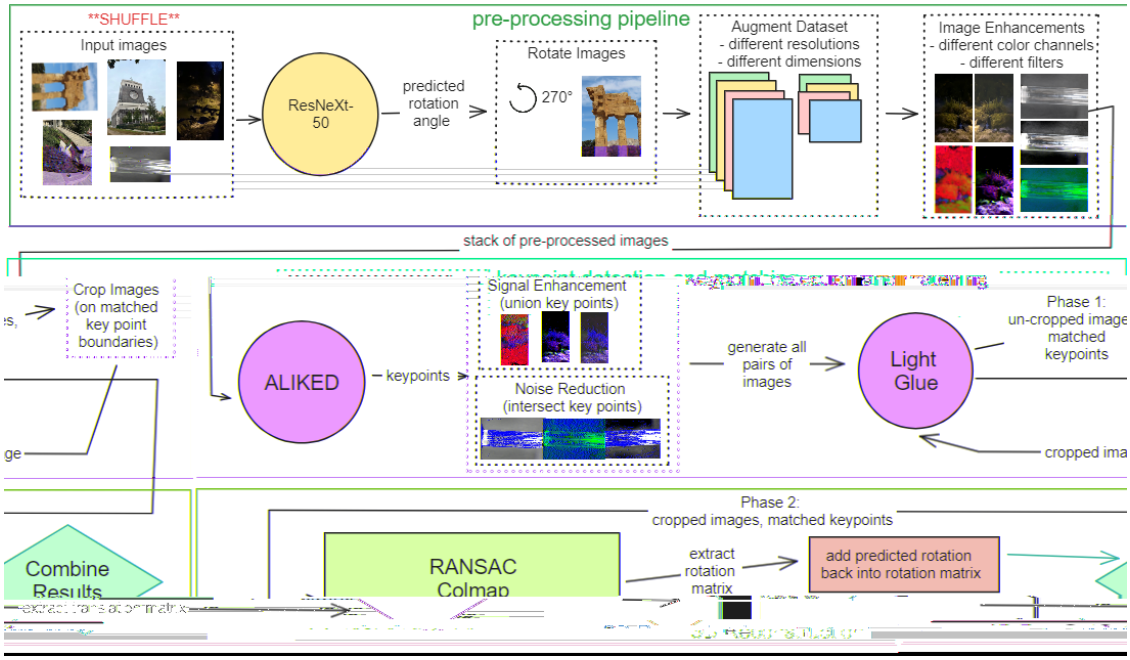
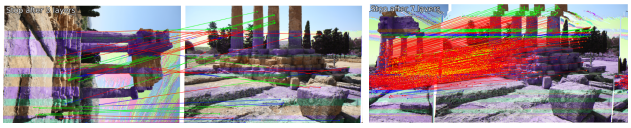


Figure 2. Proposed modifications to the SfM pipeline

turing order, we wanted to ensure that the training dataset captured a similar distribution by shuffling the input images before being ingested into the pipeline. Based on the exploratory data analysis result where the model had difficulty matching keypoints between rotated and un-rotated images, we used a pre-trained ResNext-50 model [6], to detect the rotation angle of the image, perform the rotation correction, and cache the rotation angle to be used in our final step when re-constructing our rotation matrix. After the rotation correction, since the number of provided training images was rather small, we decided to augment the dataset by adding multiple image resolutions. We also enhanced the images with a variety of transformations such as converting to different color channels (i.e. HSV, LAB, etc.), and increasing the contrast and brightness. The assumption was that for darker images, uniform images, and translucent images these enhancements would allow the model to more easily detect keypoints [5]. The goal of our pre-processing modifications is to generate images that will produce higher quality keypoint matches to feed into the 3D reconstruction phase.

Next, the stack of pre-processed images undergo our keypoint detection and matching phase where the ALIKED model, which is a newer keypoint detection model compared to SIFT, generates a set of keypoints for each image. For images that give the model difficulty such as dark images or uniform images, we perform signal enhancement where we take the union of the generated keypoints over the stack of pre-processed images. Conversely, when there is too much noise and the model generates too many keypoints, we take the intersection of the keypoints over the stack of images. The logic to decide whether we union or intersect the images relies on a set of image quality metrics we compute over the mean of the brightness, contrast, and saturation of the images. After we have the keypoints for each image, we exhaustively generate every possible image pair and pass these pairs of images along with their keypoints to the LightGlue model to match the keypoints. We split the matching phase into 2 separate phases: first we simply match the keypoints for the pair of images, then we crop the pair of images based on the a computed bounding box of the matched keypoints using minimum and maximum values for each x and y axes. The cropped images along these matched keypoints are again passed to the LightGlue model to perform keypoint matching again. We found that the keypoints are matched more accurately in this second pass which also supports our goal of aiding in the 3D reconstruction phase.

Lastly, we take the matched keypoints as well as their descriptors through a RANSAC algorithm using colmap to perform 3D reconstruction. Once the scene has been reconstructed, we extract the rotation matrix and transformation matrix for each image in the 3D reconstruction. If the image was previously rotated in our pre-processing phase, we make sure to add back the rotation angle into the rotation matrix to get our final rotation matrix.



(a) Pre-rotation, 27 matches (b) Post-rotation, 639 matches

Figure 3. Keypoint matching improvement from image rotation

Method/Model	Church	Dioscuri	Pond
ALIKED (baseline)	0.0	0.0	0.0
SuperPoint	-0.012	-0.009	+0.002
Rotation correction	-0.003	+0.157	-0.002
Image resolutions	+0.012	+0.009	+0.031

Table 1. mAA improvement deltas over baseline

4. Experimental Results and Discussion

ALIKED and LightGlue are rotation variant keypoint detection and matching models, so these models benefit from a pair of images having the same general rotation to begin with. The original pair in Figure 3a found only 27 matches in each image, most of those being incorrect, as it is expecting vertical features to remain vertical. After taking the rotation of the image into account as shown in Figure 3b, the model found a dramatic improvement of 639 matches. To account for this rotation pre-processing in determining the rotation pose of the camera, we multiplied the rotation pose outcome by the initial 3x3 rotation matrix. Applying this method on the entire dataset for two scenes delineates the results–the church scene, having no rotated images, sees no improvement from the rotation correction as expected, while the dioscuri scene, with several rotated images, sees an improvement of 0.053 (6%) on the mAA (Table 4).

After the initial SfM pipeline was established, we tried different variations of keypoint detection models such as including SuperPoint [2], SIFT [9], and DISK as well as different keypoint matching models as SuperGlue. Given the tradeoffs in both model accuracy as well as computationally efficiency, we choose ALIKED and LightGlue to be our final keypoint detector and keypoint matcher in our SfM pipeline but made sure to perform rotation correction for these models due to their sensitivity to rotated images.

To improve on the baseline keypoint matching model, we took advantage of Test-Time Augmentation (TTA). This is a heuristic where we augment the input dataset by creating variations of the original input for consideration, to ultimately produce a summed output that may contain more information than the baseline. In this case, we tried the ALIKED keypoint detection model with two different image resolutions, and concatenated both sets of keypoints, which ultimately led to a higher number of keypoints detected. This is helpful over simply increasing the maximum number of keypoints (4098), as the inference time increases more rapidly by doubling the maximum number of keypoints, than from running the same number of keypoints twice [11]. Despite the improvement in runtime, the improvement in mAA was not seen, as shown in Table 4.

Datasets such as the lizard scene and the temple scene contain images taken at night, that suffer from both low light and image blur. To address this, we tried two methods: (1) applying the pre-trained DeblurGANv2 model [8] to sharpen night images, and (2) and correcting the average pixel value to the same baseline. DeblurGANv2 did not improve the results (mAA stayed the same to 0.1%), mainly because the pre-trained network was not robust to this dataset. Correcting the average pixel value showed some improvement. For simplicity and due to comparable results, we decided to continue to more traditional feature enhancement methods that ended up being a part of the final SfM pipeline.

5. Conclusion

From the proposed SfM pipeline improvements detailed in Figure 2, rotation correction was the most successful in improving the mAA. Other methods such as concatenating the key points from multiple image resolutions helped detect more key points, but did not improve the mAA. We found that the feature engineering during the image enhancement phase did slightly improve the mAA which leads up to believe that having another neural network, a dedicated CNN, for more complex feature extraction would be beneficial to this problem.

In the future, we we would also like to compare the performance of rotation correction with a rotation-variant keypoint matching model as we have shown in this paper, against the performance of a rotation-invariant keypoint detector such as KeyNet [1], AffNet, HardNet, and SOSNet, without the need to pre-correct for rotation. It would also be interesting to not only have a sparse reconstruction of the scenes in 3D but try dense reconstruction methods as well. We feel that using a CNN for better feature extraction may be useful to include in our pipeline in the future. A segmentation model may also be useful to separate the different parts of a complex scene especially when dealing with temporal changes where the background may be inconsistent but the objects in the scene may be consistent. Lastly, to better improve our methods, we could also ensemble different models together to generate multiple key points and key point matches, then union or intersect the output depending on if we would like to reduce the noise or garner more signal in scenes where key points may be difficult to extract. Then, we could perform a hyperparameter search to find the best hyper parameters for each of our models and thresholds in the pipeline.

6. Individual Contributions

We worked together to understand the problem statement clearly, the evaluation metrics, and previous works and methods. Then, we both spent time on exploratory data analysis to gain a deeper understanding of the dataset and problem at hand. We were able to brain storm the potential difficulties of each scene to tackle this problem together. We both worked together on rotation correction especially since some models were rotation variant and made sure the rotations were computed correctly and discussed where they should be in the pipeline.

Kat brainstormed and focused on the pre-processing pipeling including feature enhancement ideas on how to better improve the images for the models to ingest, focusing on the concept of "garbage-in, garbage-out". This involved looking into different feature enhancement ideas to generate the best keypoints for the matching step. She also worked on detecting the region of overlap between the matched keypoints into multiple phases to crop the images so that the matching algorithms could have better signal to re-compute on a more focused region of the images.

Sae Na brainstormed and focused on the keypoint detection and matching models and methods. She investigated the model architectures to understand the tradeoffs and tested which models would best work within the constraints of our SfM pipeline. In addition, she also investigated different models outside of keypoint detection and matching such as the potential of GANs to improve our pipeline by improving darker images.

References

[1] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019. 4

[2] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Deep Learning for Visual SLAM Workshop*, 2018. 4

[3] Dmytro Mishkin Luca Morelli Fabio Remondino Weiwei Sun Amy Tabb Eduard Trulls Kwang Moo Yi Sohier Dane Ashley Chow Fabio Bellavia, Jiri Matas. Image matching challenge 2024 - hexathlon, 2024. 1

[4] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981. 1

[5] Marco Gaiani, Fabio Remondino, Fabrizio I. Apollonio, and Andrea Ballabeni. An advanced pre-processing pipeline to improve automated photogrammetric reconstructions of architectural scenes. *Remote Sensing*, 8(3), 2016. 3

[6] Vladimir Iglovikov. Check orientation by ternaus. https://github.com/ternaus/check_orientation, 2020. 3

[7] Sae Na Na Kat He. Code for comsw4732 final project. https://github.com/kat-he/CU-3D-Reconstruction/tree/main, 2024. 1

[8] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 4

[9] Tony Lindeberg. *Scale Invariant Feature Transform*, volume 7. 05 2012. 1, 4

[10] Penjani Nyimbili, Hande Demirel, Dursun Seker, and Turan Erden. Structure from motion (sfm) - approaches and applications. 09 2016. 1

[11] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks, 2020. 4

[12] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[13] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1