CSCE 57003 Computer Vision Final Project
Image Matching across Wide Baselines: From Paper to Practice

Simon Chaisouang
University of Arkansas
shchaiso@uark.edu

Xiaoyu Guo
University of Arkansas
xsguo@uark.edu

Venkata Kumari
University of Arkansas
vg021@uark.edu

1. Introduction

Image matching across broad baselines is a basic and difficult problem in computer vision that is addressed in this study. Finding similar points or features between two or more photos of the same scene taken in wildly disparate settings is the task at hand. A key component of many computer vision tasks, such as camera localization, Simultaneous Localization and Mapping (SLAM), image retrieval, and 3D reconstruction, is image matching. Reliable matching, however, is especially challenging when there are differences in the photos, such as different camera settings, partial occlusions, lighting conditions, or angles of view.

The problem is extremely non-trivial because of these difficulties. For example, it can be challenging to recognize shared elements when comparing features in photos shot from drastically different perspectives. Like this, variations in lighting or shadows can modify how surfaces appear, making it more difficult to match similar spots. Despite its significance, current approaches find it difficult to deliver reliable results in these challenging circumstances. Therefore, improving the resilience of computer vision systems in practical applications requires tackling wide-baseline image matching.

1.1. Problem Statement

The authors of the paper proposed a modular framework that can be used to compare a wide variety of different image matching methods and heuristics. The modular frameworks allowed for easy implementation and configuration, and can be used in evaluating future newly developed method against the ground truth given by Colmap. More details of the pipeline will be discussed in Section 2 On top of the framework, the authors also identified some limitations in existing datasets, so they also created a brand new dataset with 25 scenes and over 30k images that can be used in future research and evaluation.

1.2. Existing Works

Techniques for image matching have been thoroughly examined in the literature and can be broadly divided into two groups: contemporary deep learning-based techniques and traditional feature-based techniques.

Images have long been matched using traditional techniques like Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), and Oriented FAST and Rotated BRIEF (ORB). These techniques concentrate on local feature extraction from pictures that can withstand rotation, scaling, and some degree of affine transformations. The quality of matches is ensured by robust outlier rejection techniques like RANSAC (Random Sample Consensus), which are commonly used for matching once features are extracted. These techniques include k-d tree-based algorithms and brute-force. Despite their effectiveness, these techniques frequently falter under harsh situations, like abrupt changes in perspective or dynamic lighting.

A new class of image matching algorithms that seek to learn characteristics and matches straight from data has emerged in recent years with the introduction of deep learning. Convolutional neural networks (CNNs) are used by methods such as D2-Net, SuperPoint, and R2D2 to compute descriptors and identify keypoints, frequently yielding better results on common benchmarks. The usefulness of these techniques in various real-world situations is still up for dispute, though. Large volumes of training data and substantial computational resources are frequently needed for deep learning-based methods, which may restrict their use in environments with limited resources.

1.3. Proposed Solution

The study suggests a thorough benchmark for assessing image matching algorithms in order to overcome the shortcomings of current techniques and datasets. The main concept is to provide a platform that offers a reliable and practical evaluation environment in addition to enabling the integration of both traditional and contemporary approaches.

The following innovations help achieve this:

- To test and compare various combinations of feature extraction, matching, and pose estimation techniques, the authors first present a modular benchmarking pipeline. Flexibility is ensured by this modularity, which enables the evaluation of deep learning-based and conventional algorithms using the same framework. The pipeline is made to encourage thorough and methodical testing, which facilitates determining the advantages and disadvantages of any strategy.
- Second, the study presents a brand-new dataset named PhotoTourism, which consists of pictures of 25 famous sites taken in a variety of settings. The dataset is extremely demanding and appropriate for testing image matching algorithms because of the notable variation in views, illumination, and occlusion in these photographs. Additionally, the dataset contains depth maps and ground truth camera poses, allowing for a precise assessment of pose estimate accuracy.
- Lastly, rather than using intermediate criteria like repeatability or descriptor matching accuracy, this benchmark bases its evaluation on downstream performance parameters like camera pose accuracy. The suggested benchmark guarantees that assessments accurately represent the practical usefulness of the evaluated algorithms by concentrating on the final task-level results.

1.4. Contributions

It is anticipated that the paper's several noteworthy contributions will improve the status of image matching research. The development of the PhotoTourism dataset, which was created especially to solve the shortcomings of earlier benchmarks, is one of its main contributions. In contrast to previous datasets, PhotoTourism offers a wide range of difficult scenarios with precise ground truth annotations, which makes it a priceless tool for testing algorithms in practical settings.

The creation of a benchmarking system that combines traditional and contemporary methods is another significant accomplishment. In addition to being adaptable and modular, this framework promotes equitable comparisons by standardizing the assessment procedure.

The paper's empirical insights into the performance of both classical and modern methods are arguably one of its most unexpected contributions. The findings demonstrate that, in some situations, well-tuned classical techniques like SIFT or ORB can beat cutting-edge deep learning-based algorithms, despite what is commonly believed. This discovery casts doubt on the notion that contemporary deep learning approaches are intrinsically better and emphasizes the significance of reexamining and refining traditional methods.

2. Pipeline

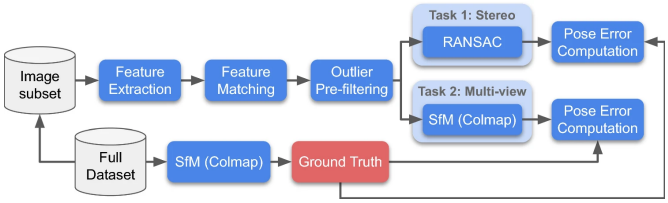
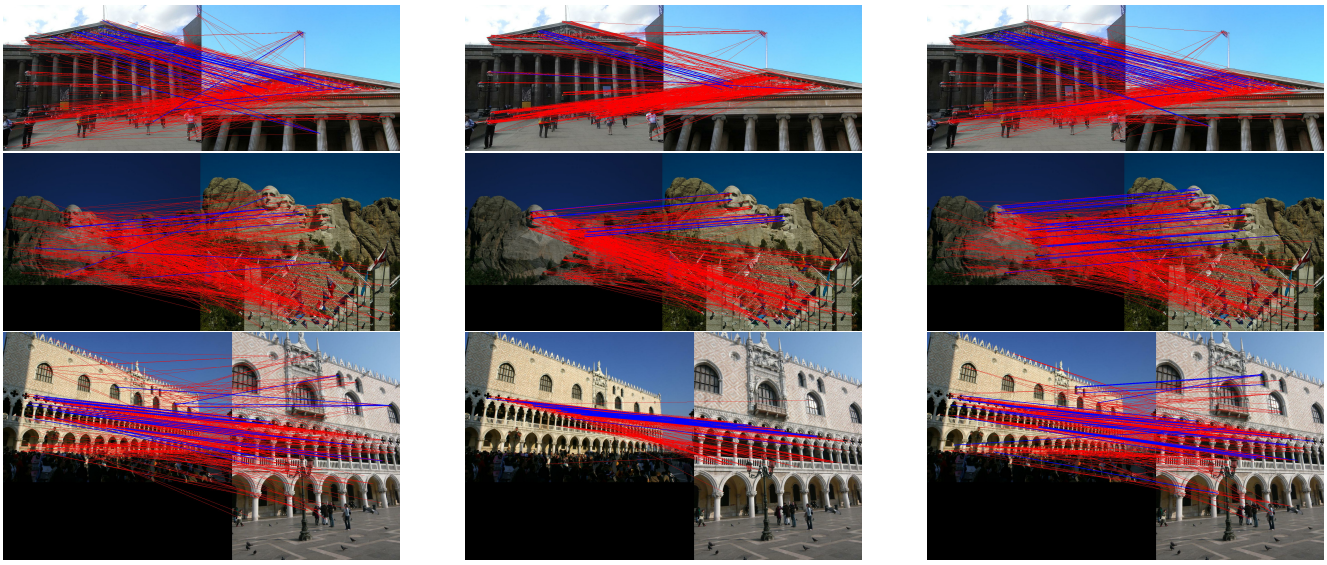


Figure 1. Proposed pipeline for image method evaluation.

Figure 1 is the framework the author proposed to evaluate different feature extraction and feature matching methods. The authors mainly focuses on the downstream task, namely stereo reconstruction and Multi-view reconstruction. For the stereo task, RANSAC will output the fundamental matrix along with the filtered matches, as RANSAC will detect the inlier from the matches from the previous steps. For the multi-view task, Colmap will be used to produce a 3D reconstruction based on the features matched. Once the outputs of the tasks are received, then we can evaluate the methods used in the pipeline against the ground truth value given by Colmap applied to the full dataset.

2.1. Fundamental Matrix

The fundamental matrix is an important concept in computer vision, especially in stereo vision and epipolar geometry. It encapsulates the geometric relationship between two images of the same scene taken from different viewpoints. The fundamental matrix, denoted as F, takes the constraints between corresponding points in two images, and describe how a point in one image relates to a line (epipolar line) in the other image. This matrix allows for the projection of points in one image to their corresponding epipolar lines in the other image, and vice versa. The fundamental matrix can be derived from corresponding points in two images using methods such as the eight-point algorithm, which requires at least eight-point correspondence. This process solves a system of linear equations derived from the epipolar constraint, generally using singular value decomposition (SVD) to ensure that the matrix satisfies the rank-2 constraint. This means that it contains two non-zero singular values. The fundamental matrix plays a vital role in application tasks such as 3D reconstruction, camera calibration, and stereo matching. In the proposed architecture, the fundamental matrix is computed by first detecting and matching features between two images. The corresponding points from each pair of images are then normalized and the F matrix is then estimated using methods like the mentioned eight-point algorithm which relates points across two images. From this information, we can obtain the camera poses to perform the 3D reconstruction of the scene which is crucial in the structure of motion (SfM).



(a) SURF (b) ORB (c) AKAZE

Figure 2. Qualitive sample results for the stereo task after applying feature detection algorithms (Blue highlights "good matches" and red highlights "bad matches").

3. Experiments

3.1. Dataset

The dataset used for this study contains images in the format of JPG files of popular landmarks and tourist attractions around the world. The dataset is divided into subsets of 100 images of the same landmark for each set. Each image is typically captured by different individuals at different times, which result in variations in lighting weather conditions, and the presence of transient objects such as pedestrians or vehicles. This data set challenges the algorithms studied in this experiment to handle real-world complexities like variation in illumination and dynamic occlusions. This makes the diversity within each subset of images valuable for further research in computer vision and 3D reconstruction.

3.2. Data Processing

A file key in the format of a text file that contains a list of all the image titles pertaining to a given subset was created in order for the source code to read in each of the raw jpg image files to be used in our recreation of the pipeline.

3.3. Experimental Results

For this experiment we focused on testing and analyzing three classical feature detection algorithms: AKAZE, ORB, and SURF. We refer to Figure 2 as sample outputs after doing feature matching with the previous pairs of images.

From the results we can see that the AKAZE algorithm was able to showcase more efficient results in terms of identifying more good matches while maintaining accuracy. This is to be expected since the algorithm is designed to be performance-efficient while maintaining sufficient computation speed. Runtime for the ORB algorithm took notably a shorter time compared to the others, but did fall short in terms of the number of good matches that were identified. The SURF algorithm performed poorly than expected compared to the other two. Although the algorithm is known to be very robustness, there are more notable mismatches between the pair of images even though they were returned as "good matches" according to our network architecture. After performing image matching across the given 50 pairs of images across each subset the RANSAC algorithm was applied to identify inliers and compute the fundamental matrix. At this step for all algorithms, almost the same number of about 24-25 pairs of images couldn't proceed to this step due to having an inefficient number of good matches identified.

References

[1] Yaqing Ding, Václav Vávra, Snehal Bhayani, Qianliang Wu, Jian Yang, and Zuzana Kukelova. Fundamental matrix estimation using relative depths. In *European Conference on Computer Vision*, pages 142–159. Springer, 2025.

[2] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021.