Feature Matching:
Using Deep-Learning Based SuperGlue

Daniel Scott and Rencheng Wu
University of Arkansas
Department of Mechanical Engineering
Date: May 9, 2024

1 Introduction

In the realm of computer vision and robotics, feature matching serves as a cornerstone for various applications, such as simultaneous localization and mapping (SLAM), object recognition, and real-time tracking. Feature matching involves identifying corresponding keypoints and descriptors between different images or frames of a scene, facilitating tasks such as 3D reconstruction, navigation, and augmented reality. This process has traditionally been completed by using handcrafted algorithms, however, deep-learning based methods have gained massive popularity within recent years.

One of the fundamental challenges in feature matching lies within the robustness to changes in viewpoint, lighting conditions, occlusions, and scale variations. Traditional methods often struggle to maintain accuracy and efficiency across diverse scenarios. Challenges with traditional methods surface during the construction of keypoints and descriptors, primarily stemming from sub-optimal weights used during the computational process. These issues prompt the exploration of advanced techniques, particularly those rooted in deep learning.

In recent years, the advent of deep learning has revolutionized the field of computer vision, offering unprecedented capabilities in feature extraction, representation, and matching. Among these innovations, the SuperGlue algorithm has emerged as a promising middle-end solution, leveraging deep neural networks to perform robust and efficient keypoint and descriptor optimization across a wide range of environments and conditions.

This report explores the concept of feature matching using the SuperGlue method, delving into its principles, applications, and significance in the domain of computer vision and robotics. We aim to elucidate the role of SuperGlue in advancing key applications such as SLAM, where accurate and real-time feature matching is paramount for precise localization and mapping in dynamic environments.

2 Literature Review

The fundamental objective in computer vision and image analysis is to identify features (distinctive points or landmarks) within images. This facilitates the establishment of correspondences or links between different views and frames within and across image datasets. Such capabilities are crucial for various advanced applications including object recognition, image stitching, and Simultaneous Localization and Mapping (SLAM), enhancing both the robustness and accuracy of spatial understanding and navigation systems [1]. These applications are widely needed in various fields, but need to first be explained.

Object Recognition: By recognizing specific patterns or features, systems can classify and differentiate between objects within diverse environments. Image Stitching: This process involves merging multiple photographic images with overlapping fields of view to produce a segmented panorama or high-resolution image sequence. Simultaneous Localization and Mapping (SLAM): SLAM techniques are essential for robotic mapping and navigation, particularly in unknown environments, where the challenge is to build or update a map of an unknown environment while simultaneously keeping track of the agent's location [1].

Keypoint detection and feature matching techniques are broadly classified into two categories: Traditional Feature Matching and Deep Learning-Based Feature Matching. Traditional methods utilize handcrafted algorithms to detect interest points (keypoints) and describe their local features. Notable traditional methods are: SIFT (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features), and ORB (Oriented FAST and Rotated BRIEF). SIFT is recognized for its robustness in detecting and describing local features in images, SIFT features are invariant to image scale and rotation, and provide robust matching across a substantial range of affine distortion, addition of noise, and changes in illumination [2]. SURF is an enhanced version of SIFT that is faster to compute and more efficient, which is crucial for real-time applications. ORB improves on both speed and performance by building on FAST keypoint detection and BRIEF descriptors.

Deep-Learning Based Feature Matching: These methods employ neural networks to automatically learn feature representations directly from the data, which can lead to more robust features under varying conditions. A few deep-learning based methods are DeepDesc, LIFT (Local Invariant Feature Transform), and SuperPoint. DeepDesc is a deep learning approach to learn descriptors for image patches, facilitating better matching accuracy. LIFT builds upon the concept of SIFT but integrates deep learning to adaptively learn from the data, by updating the weights. SuperPoint: is a neural network trained in a self-supervised manner to detect interest points and describe their attributes simultaneously. It has demonstrated substantial improvements in tasks requiring feature matching across different scenes [3].

Recent advancements, especially in deep learning-based methods, are setting new benchmarks in feature detection and matching. The development of SuperGlue, a model that uses Graph Neural Networks, exemplifies the trend towards more adaptable and robust feature matching techniques [4]. These innovations are pivotal for real-time applications and systems requiring high levels of perceptual accuracy, from autonomous vehicles to augmented reality systems.

3 Methodology

SuperGlue is a keypoint and descriptor optimization method that can be used in feature matching applications. The architecture of SuperGlue consists of two main component, a graph neural network and an optimal matching layer, these components can be seen in Figure 1. SuperGlue will take keypoints and descriptors from a separate method (in our case SuperPoint), and will encode the keypoints and descriptors to then develop a self and cross graph. The self graph connects keypoints together within the same image, building a map for each specific keypoint. The cross graph will then attach the keypoints between the graphs. This can then be sent into the optimal matching layer, which will solve the optimization problem, using the sinkhorn algorithm.

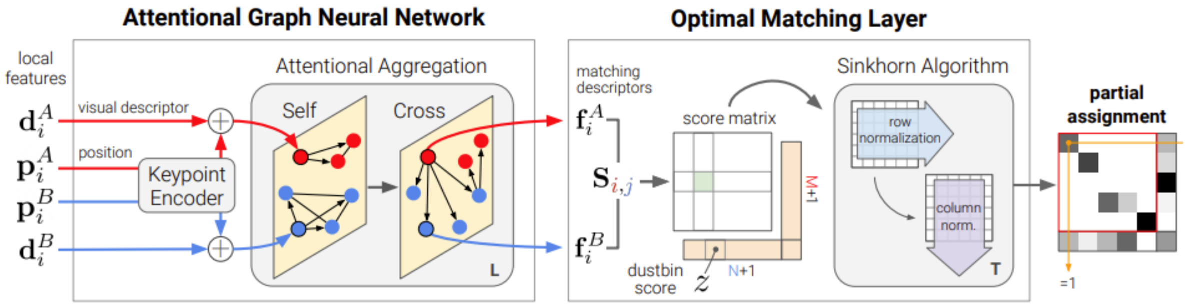


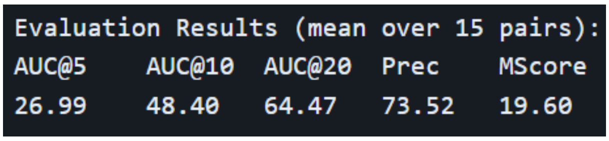
Figure 1: SuperGlue main architectural components

The optimization problem will provide an understanding of the keypoints that can potentially be matched together. This is based on both the self and cross graphs, that allow the network to perform with more detailed information. This information provided gives the overall landscape around the keypoint, which can assist in understanding the matches. SuperGlue does not need initial weights, like traditional methods, to run. Therefore, the training of SuperGlue consists of inputting pairs of images, with the ground truth values.

Training SuperGlue also depends on the trained keypoint and descriptor method used. As previously stated, SuperPoint is the keypoint and descriptor method utilized in this work. SuperPoint is a deep-learning based feature matching method that has shown promising results, when compared with traditional methods. SuperPoint is first trained on the input dataset, to then be utilized train SuperGlue. In both cases, the weights within the networks is updated until the optimal results are found. These results were then saved and provided with test code.

4 Experimental Results

Performing feature matching utilizing SuperGlue, there is a requirement for a keypoint and descriptor development method. In literature, SuperPoint in combination with SuperGlue produced the overall best results, so we set out to attempt to recreate these results, and determine potential avenues to increase the accuracy. The provided code from [5] gives all the needed code and data to reproduce the results found in the paper. After rewriting the code, we were able to produce the same results as [4], which can be seen in Figure 2.



```
Evaluation Results (mean over 15 pairs):
AUC@5    AUC@10  AUC@20  Prec    MScore
26.99    48.40   64.47   73.52   19.60
```

Figure 2: Results produced from the pre-trained model using the provided scannet dataset

Now that the results produced match the literature results, we utilized SuperGlue to match features within the AM3 Robotics Lab Platform. This platform consists of 7 robots arranged in a random assortment within the lab setting. We attempted to test the feature matching abilities and the real-time capabilities, the results can be seen in Figure 3. The results show the accuracy with the produced methods, across a variety of datasets.

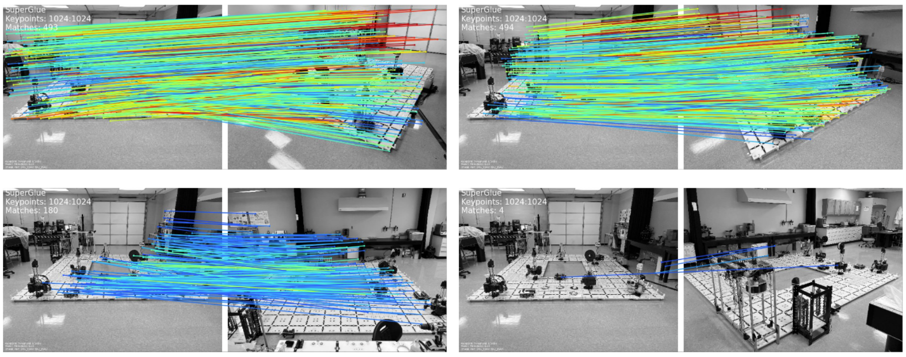


Figure 3: AM3 Lab Robotics Platform Results

Now that we produced the same results and tested our own dataset, we attempted to retrain the model with our new datasets (Phototourism [6] and YFCC [7]). However, in researching the provided literature and discussions, regarding SuperGlue and SuperPoint, the training code and methods were not provided. Due to the time constraints, we attempted to update the weights within the pre-trained models to then test with our results. This method proved promising providing effective results for the Phototourism dataset. After updating the weights we were able to produce results such as Figure 4.

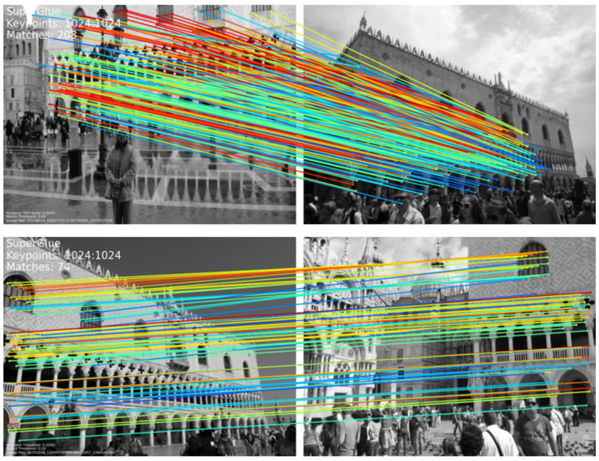


Figure 4: Phototourism results after updating models

Once we received these promising results, we attempted to utilize the YFCC100M dataset [7], to determine the effectiveness of the updated weights. The YFCC100M dataset provided many errors, however. The dataset is very large, so we determined to only download a portion (images from 000 to 199, which took up approximately 500GB). We received further issue regarding the dataset, due to the computer freezing after download, therefore, half of the dataset was deleted. This posed a problem, since the current YFCC dataset did not contain the file discussing the pairs and ground truth, we attempted to develop our own. This lead us to creating a code to generate the pairs, which can be seen in ”Pairs Generator.py”, which loaded the images and created the pairs in the dataset.

Now that we have the pairs, we needed to match the ground truth values we have, so another code was created to update the indexes to match the pairs to the ground truth values, which can be seen in ”Update index for pairs.py”. This process, however, proved to develop issues within the images and the model. Therefore, the results were not fully generated correctly, providing us with the .npz files, but not the visualization or the evaluation results. If further time was alloted, the pairs of images would be refined and reformated to work within SuperGlue to test the updated models.

5 Discussion

SuperGlue is a promising method to optimizing keypoints and descriptors within a dataset. SuperGlue has show overall better results when compared to traditional and existing deep-learning based feature matching methods. After producing the SuperGlue and SuperPoint code, we tested to ensure the given results are true and have shown the results to be correct. After this we utilized our own dataset to understand the generalizability of SuperGlue in its current state, which showed an effective production of new results. After this, we attempted to improve SuperGlue by updating the weights, which showed more promising results.

Overall, SuperGlue is an effective feature matching method, producing highly accurate results. However, during our testing, it became apparent that SuperGlue is heavily dependent on the keypoint and descriptor development method. Therefore, to further improve upon SuperGlue, we should develop a keypoint and descriptor method within the framework of SuperGlue or improve the overall results from existing methods.

References

[1] H. F. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part i," *IEEE Robotics & Automation Magazine*, vol. 13, pp. 99–110, 2006. [Online]. Available: `https://api.semanticscholar.org/CorpusID:8061430`.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004. [Online]. Available: `https://api.semanticscholar.org/CorpusID:174065`.

[3] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 337–33 712, 2017. [Online]. Available: `https://api.semanticscholar.org/CorpusID:4918026`.

[4] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," *CoRR*, vol. abs/1911.11763, 2019. arXiv: `1911.11763`. [Online]. Available: `http://arxiv.org/abs/1911.11763`.

[5] *SuperGluePretrainedNetwork: SuperGlue: Learning feature matching with graph neural networks (CVPR 2020, oral)*, en.

[6] *IMC-PT 2020 dataset*, `https://www.cs.ubc.ca/~kmyi/imw2020/data.html`, Accessed: 2024-5-9.

[7] *The multimedia commons initiative*, en, `https://multimediacommons.wordpress.com`, Accessed: 2024-5-8.