Comparative study between DeLF and SIFT for Landmark Detection

Maddali Sowmya
Dept. of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India-560098
msowmya1ga17cs080@gmail.com

Swaraj Parida
Dept. of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India-560098
swaraj1ga17cs163@gmail.com

C.P Yashwanth
Dept. of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India-560098
cpyashwanth1ga17cs035@gmail.com

Rakshitha Murthy
Dept. of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India-560098
rakshitha1ga17cs118@gmail.com

Kamleshwar Kumar Yadav
Dept. of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India-560098
kewalkamlesh07@gat.ac.in

***Abstract*— The world of Computer Vision is vast and complex. This world can bring about problems that is resilient to scale along with maintaining the accuracy. Identifying landmarks in large-scale is problematic. Number of landmarks in a dataset can be sky-high and can be still scalable, so identifying one landmark out of a bulk will cost us high computational resources and time. Thus, the approach of this paper is to do a comparative study between two different models to ensure how they individually and significantly impact on the results. Existing datasets are either enormous or lack variety of landmarks. Consequently, we are introducing a dataset which is coherent amongst different categories and is not computationally heavy to use.**

Keywords—Computer Vision, SIFT, DeLF, VGG

I. INTRODUCTION

Large-scale recognition is the main concern and our paper deals with the variable database size and can highly scalable. Also, the objects that we are detecting are at an instance level so, identifying features that defines the instances is the key.

Present datasets [1] are tremendous which makes it difficult to be scaled, therefore require high computational resources and time whereas our dataset performs well while having vast variety and keeping all the other factors to a minimum.

Simple CNN architectures [5] are not built explicitly for capturing important features of landmarks and they produce extremely limited accuracy in reality. So, we test the dataset on a simple approach of identifying feature vectors of the images and using it to perform a similarity search to fasten the retrieving process. For feature matching Scale Invariant Feature Transform (SIFT) extracts the keypoints and computes its descriptors. Then keypoint matching is used to find inliers between query image and similar images. The result of the approach is compared with state-of-the-art model.

Recognition is done using Deep Neural Network with CNN layers and trying to extract features of an image.

For retrieving information from a large-scale dataset, we are comparing feature vector of query image with all the nearest neighbor feature vectors from the dataset. This is done with respect to two different models namely: SIFT [3] and DeLF [2].

SIFT is a method for recognizing important, stable feature points in a photograph. It also gives a set of "features" for each such point. These characteristics are unaffected by rotation or scale.

DEep Local Feature (DeLF) module, which is available on TensorFlow Hub, can be used to substitute alternative keypoint detectors and descriptors for image retrieval. It uses feature descriptors, which are 40-dimensional vectors that describe each notable point in an image.

II. DATASET

The inspiration for making a new dataset comes from the obstacle we have faced from the pre-existing datasets available. Currently, the dataset which consists of landmarks, are either extremely large and computationally heavy to use or they are not diverse enough to help identify any given landmark.

Our dataset has 2000 images consisting of 200 landmarks. Knowing that the dataset is small to capture all the landmarks we are still trying to diversify based on classification of landmarks. The landmarks are classified as Monuments, Buildings, Towers, Religious Monuments Palaces, Bridges, Dams, Statues, Mountains, and Waterfalls. These categories ensure that there is variety in the dataset and any landmark that comes under it can be identified. A small dataset is easy to handle, store, process, and beats the computational power required to work on the present state-of-the-art dataset.

Figure 1: A glimpse at our custom dataset

Annotations is a costly technique on a large dataset that can be scaled up. So, using annotations on a scalable dataset is not feasible as annotating a large dataset takes a lot of time and is computationally heavy. Since all images in the datasets are landmark-centric, which helps global feature description work well.

The dataset made is an example of anyone trying to apply landmark recognition to their problem.

III. APPROACHES FOR LARGE-SCALE RECOGNITION

The idea of Feature Extraction is to represent each image in form a feature vector. Since working on images is a computational problem and we cannot take each landmark as a class as this problem is scalable. When the query image is in the form of vector which is a linear data rather than pixels which helps in easier comparison. We get the similar images from the database and perform feature matching. The image with the highest number of inliers is returned.

A. Feature extraction using Visual Geometric Group – 16

A convolutional layer is followed by one or more dense (or fully connected) layers in the VGG16 model. This network is quite huge, with approximately 138 million (approx.) parameters. The feature extraction part of the model runs from the input layer to the last max pooling layer (labelled by 7 x 7 x 512), whereas the other layers are the classification part. We need to load the input image with the size that the model expects after we have defined the model. The output of the maxpooling layer is used to extract features. The feature vector has a shape of (1,7,7,512), or 25088 dimensions.

Because the obtained feature vectors are significantly larger per image, the output layer of the VGG16 is processed through an autoencoder pipeline, which is made up of two parts: an encoder and a decoder, both of which are dense neural networks with batch normalization and relu activation. The encoder is used to compress the feature vector, reducing the number of superfluous feature vectors while increasing the number of notable feature vectors.

B. Image Recognition using DeLF module

DeLF [2] was trained on [1] over a million images containing landmarks and additional query images optimized for landmark recognition. Image recognition using DeLF approach is consists of four basic components: “Dense localized feature extraction, Keypoint selection, Dimensionality reduction, Indexing and retrieval. The first 3 blocks are wrapped into the TensorFlow Hub DeLF module. Even so, it's still interesting to crack open the black box and look inside”.

They used a two-stage training technique, in the first stage ResNet50 layers are fine-tuned to improve the local descriptors [4], then using the attention module to measure the importance of the features extracted by the model. Here the local descriptors will be used to match against the query image.

The feature dimensions are decreased to 40 via PCA in the dimensionality reduction step, while maintaining the trade-off between compactness and discriminativeness. This reduces the amount of memory required to retain the features of all photos in the database while still allowing similar images of the query image to be found.

IV. RETRIEVAL USING INDEXING

A. Retrieval using SIFT

SIFT [3] identifies as keypoints and saves them in a database. These keypoints are quantified as descriptors that can be used for feature matching to perform the object recognition task. When a query image contains an object, it is recognized by comparing its features to the database of keypoints and descriptors. The comparison is done based on the Euclidean distance between the feature vectors.

SIFT uses the scale Invariant feature detection technique. This method converts a picture into a huge number of feature vectors. These characteristics are resilient to local geometric distortion and invariant to picture translation, scaling, and rotation. They are also largely invariant to illumination variations.

- Once potential keypoint locations have been identified, they must be enhanced in order to provide more accurate findings. To acquire a more precise location of extrema, they used Taylor series expansion of scale space.
- To achieve image rotation invariance, each keypoint is given an orientation. Depending on the scale, a neighborhood is drawn around the keypoint location, and the gradient magnitude and direction are determined in that area.
- A keypoint descriptor is then established. To form a keypoint descriptor, it is represented as a vector.
- The closest neighbors of keypoints in two images are identified and matched. It uses the best-bin-first search method. This is a variation of the k-d tree algorithm that can identify the nearest neighbors with high probability.

B. Retrieval using DeLF

Retrieval begins by extracting feature descriptors from the landmark picture database in the recognition step. Their descriptors and positions are combined. A KD-tree is built which aggregates descriptors to help with our picture retrieval method, which is based on closest neighbor search. Unless we want to add more images into the database in the future, the KD-tree is built only once and offline indexing is done. At runtime, the DeLF module computes the query image's descriptors and locations after resizing and cropping it to 256x256 resolution. Then, for each descriptor of the tree, we call the KD-tree to discover K nearest neighbors. Next, collect all the matches for every database image. Finally, we use RANSAC method to perform geometric verification and the score for the retrieved images is based on the number of inliers. When using RANSAC for geometric verification, it should be kept in mind that all the matches should be consistent with a global geometric change.

RESULTS

a) SIFT approach results

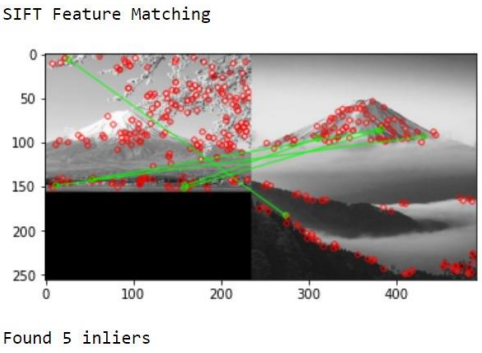

Figure 2: Landmark considered here is Mount Fuji and feature matching is done between the same landmarks which results in 5 inliers. The number of inliers is less and not all inliers are accurate.

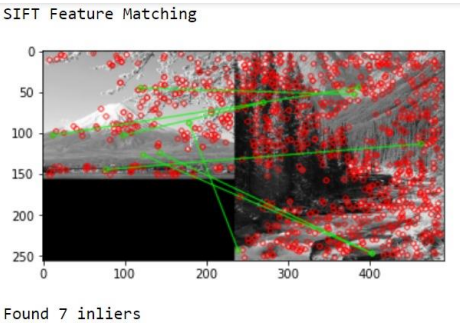

Figure 3: Landmarks considered here is Mount Fuji and a Forest landmark for which feature matching is done and it results in 7 inliers. The background details such as trees are being matched.

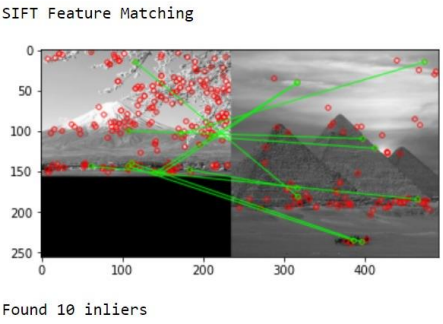

Figure 4: Landmarks considered here is Mount Fuji and The Great Pyramid of Giza for which feature matching is done and it results in 10 inliers even though the two landmarks are vastly different.

Challenges faced:

- Since SIFT uses the global features to identify the landmark other background information also gets included in the matching process.
- The background becomes noise in the feature vector and hence the results of instance-based recognition is defective to use.

b) DeLF approach results

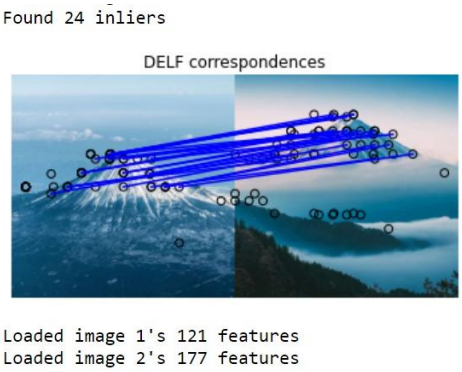

Figure 5: Landmark considered here is Mount Fuji and feature matching is done between the same landmarks which results in 24 inliers. The number of inliers is higher than other matches and is accurate for practical use.

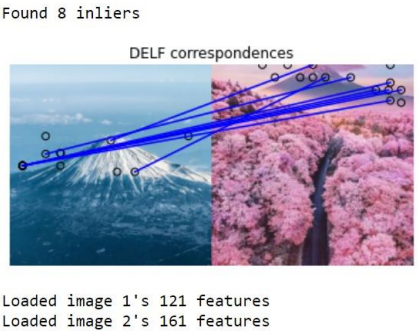

Figure 6: Landmarks considered here is Mount Fuji and Mount Fuji during a different season for which feature matching is done. Although the background and a lot of area in the image is entirely different still local features extraction in DeLF helps identify 8 inliers in such a corner case.

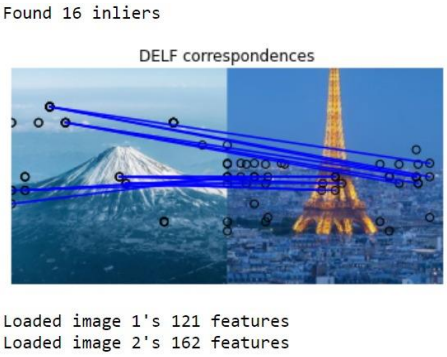

Figure 7: Landmarks considered here is Mount Fuji and Eiffel Tower for which feature matching is done and it results in 16 inliers. In some scenarios like this, the matches do happen with some noisy features, but DeLF is powerful enough to provide higher number of inliers with the correct landmark.

Challenges faced:

- Images with multiple landmarks can hinder the feature matching process.
- If the landmark has been occluded by an object, then DeLF can face difficulty in matching inliers.

CONCLUSION

The present dataset is very large and difficult to perform comparisons on. Our paper is based on Comparison of different models, SIFT and DeLF, on custom dataset which is relatively smaller than the present datasets. The approach of DeLF is based on identifying the local features which is more accurate than SIFT which identifies only the global features . We have compared images with noise and found that DeLF gives more accurate results.

REFERENCES

[1] André Araujo, Bingyi Cao, Jack Sim, Tobias Weyand, "Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retieval," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2572-2581.

[2] André Araujo, Bohyung Han, Hyeonwoo Noh, Jack Sim, Tobias Weyand, "A Large-Scale Image Retrieval with Attentative Deep Local Features," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp.3456-3465.

[3] Henna R. Kher, Vishvjit K. Thakar, "Scale Invariant Feature Transform Based Image Matching and Registration," 2014 Fifth International Conference on Signal and Image Processing, 2014, pp. 50-55.

[4] Jeff Donahue, Jitendra Malik, Ross Girshick, Trevor Darrell, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587.

[5] Ross Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448.