Task-5 DJI Datasets and Feature Detection & Matching Algorithms

Hemanth Balaji Dandi
hemanthdandi.aero2astro@gmail.com
21 May 2021

1 DJI Dataset(s) Description and Pre-Processing

Few major Pre-Processing techniques include Color-conversions, Resizing, Normalization, Maybe PCA (For Dimensionality reduction of multi-spectral images), Augmentation (Deep Learning), Edge detection (During Segmentation), DeBlurring/Denoising.
A typical Drone dataset also has meta-data which provides information about camera parameters required for calibration, GCPs, CPs, (If any) to improve accuracy, GSD , coverage, flight altitude, number of images, and a Flight trajectory file. (Source: Looking at Sensefly)
Some pre-processing done on Drone Aerial images (Sensefly) along Image Stitching will be shown in the notebook.

2 Image Stitching

- Extract key-points and descriptors from each image
- Perform Feature Matching (Brute-Force/Flann-based) by identifying good candidate pairs (RANSAC)
- Perform Perspective Warping of the images together by computing the Homography matrix to warp one image with the other image.

3 Notebook for Image Stitching

https://colab.research.google.com/drive/1fW53YCYLg4oB30iSsxHPuVyxJz2Wu_eu?authuser=2#scrollTo=8ZlzW4P6YROw

4 Traditional Feature Detection and Matching Algorithms

4.1 Different Feature Keypoints and Descriptors

Feature Keypoints refers to the important high-frequency information of the image like edges,corners,etc. Feature Descriptors are matrix of descriptions/characteristics associated with each keypoint, and should be scale,rotation,brightness invariant to actually capture the correct information of the objects present within the image.

Below are some Algorithms used for extraction keypoints and descriptors:

4.1.1 SIFT

- Scale Invariant Feature Transform

- Consists of both KeyPoint and Descriptor
- Uniform scaling, Orientation and Brightness changes invariant, partially invariant to Affine Distortion
- Not free for commercial use

SIFT consists of four major stages: scale-space extrema detection, keypoint localization, orientation assignment and keypoint descriptor. The first stage used difference-of-Gaussian function to identify potential interest points , which were invariant to scale and orientation. DOG was used instead of Gaussian to improve the computation speed In the keypoint localization step, they rejected the low contrast points and eliminated the edge response. Hessian matrix was used to compute the principal curvatures and eliminate the keypoints that have a ratio between the principal curvatures greater than the ratio. An orientation histogram was formed from the gradient orientations of sample points within a region around the keypoint in order to get an orientation assignment . According to the paper's experiments, the best results were achieved with a 4 x 4 array of histograms with 8 orientation bins in each. So the descriptor of SIFT that was used is 4 x 4 x 8 = 128 dimensions.

4.1.2 SURF

- Speeded-Up Robust Features
- Consists of both KeyPoint and Descriptor
- Similar to SIFT but many times faster
- Patented

SIFT and SURF algorithms employ slightly different ways of detecting features. SIFT builds an image pyramids, filtering each layer with Gaussians of increasing sigma values and taking the difference. On the other hand, SURF creates a "stack" without 2:1 down sampling for higher levels in the pyramid resulting in images of the same resolution. Due to the use of integral images, SURF filters the stack using a box filter approximation of second-order Gaussian partial derivatives, since integral images allow the computation of rectangular box filters in near constant time.

In keypoint matching step, the nearest neighbor is defined as the keypoint with minimum Euclidean distance for the invariant descriptor vector. Lowe used a more effective measurement that obtained by comparing the distance of the closest neighbor to that second-closest neighbor so the author of this paper decided to choose 0.5 as distance ratio like Lowe did in SIFT.

4.1.3 FAST

- Features from Accelerated Segment Test
- Has only Keypoints
- Not robust to high noise and depends on a threshold
- Not rotationally invariant

FAST is based on a characteristic feature criterion Accelerated Segment Test (AST). This considers a circle of 16 pixels constituting a discrete circle around the center-pixel. It compares each pixel's intensity on a circle with the center pixel P. If there exist more than S connected pixels on the circle with intensities greater than P's intensity plus a threshold, T, or all of them less than P's intensity minus a threshold T, the center pixel is considered a feature. T is a user defined threshold

4.1.4 BRIEF

Binary Robust Independent Elementary Features (BRIEF) uses binary strings as an efficient feature point descriptor, which is highly discriminative even when using relatively few bits and can be computed using simple intensity difference tests. Furthermore, the descriptor similarity can be evaluated using the Hamming distance, which is very efficient to compute, instead of the L2 norm as is usually done, thereby making BRIEF very fast to build and to match.

4.1.5 ORB

The ORB algorithm combines the improved corner detection operator (FAST) and the BRIEF descriptor, and its calculation speed is improved considerably; however, it lacks scale invariance.

4.1.6 KAZE

KAZE detects two-dimensional feature points by non-linear diffusion filtering in the non-linear scale space; it preserves the boundary of the object and achieves high accuracy and particularity. However, it involves high computational costs.

4.1.7 AKAZE

The accelerated version of the KAZE algorithm (AKAZE) uses the fast explicit diffusion (FED) numerical analysis framework to solve the non-linear diffusion filtering equation, which enhances the accuracy and reduces the implementation complexity

4.1.8 SRP-AKAZE

Sparse Random Projection-AKAZE (SRP-AKAZE) follows the following 4 steps for fast multiscale feature detection and description: Construction of a non-linear scale space by a non-linear filtering function that can be solved using the FED algorithm; Feature detection using the normalised Hessian matrix with different scales, Identification of the dominant orientation using SIFT descriptor based on SRP; and feature matching is determined by calculating the similarity of the feature description vectors of two images using nearest neighbours ratio test.

4.1.9 BRISK

- Binary Robust Invariant Scalable keypoints
- It is an invariant version of AGAST in multiple scale spaces
- Has a low repeatability, which can cause drift issues with feature tracking

The BRISK algorithm extracts corner features through the adaptive and generic accelerated segment test, and it constructs binary descriptors based on pixel grey value comparison. Although its calculation speed is high, its robustness is poor.

4.1.10 Multi-Resolution MSER

Maximally Stable Extremal Regions (MSER) is a method for blob detection in images. The MSER algorithm extracts from an image a number of co-variant regions, called MSERs: an MSER is a stable connected component of some gray-level sets of the image . It follows the following stages-Instead of detecting features only in the input image, construct a scale pyramid with one octave between scales,Detect MSERs separately at each resolution

.Duplicate MSERs are removed by eliminating fine scale MSERs with similar locations and sizes as MSERs detected at the next coarser scale.

Multi-resolution MSER provides better robustness to large scale changes and blurred images than MSER and improves matching performance over large scale changes and for blurred images. It good repeatability, affine invariant, and one of the fastest feature detectors.

4.1.11 AGAST

Adaptive and Generic Corner Detection Based on the Accelerated Segment Test feature detector (AGAST) is trained based on a dataset that includes all possible combinations of 16pixels on the circle. This ensures that the decision tree works in whatever environments. Moreover, AGAST introduces a dynamic tree switching algorithm, which automatically changes the decision trees. One tree is trained under homogeneous areas, and the other is trained under heterogeneous areas. In this way, the performance of AGAST increases for random scenes. By combining these two improvements, It works in any arbitrary environments without any training steps. It is npt rotationally invariant

4.1.12 LDB

Local Difference Binary (LDB) is a highly efficient, robust and distinctive binary descriptor. The distinctiveness and robustness of LDB are achieved through 3 steps. First, LDB captures the internal patterns of each image patch through a set of binary tests, each of which compares the average intensity Iavg and first-order gradients, dx and dy, of a pair of image grids within the patch. Second, LDB employs a multiple gridding strategy to capture the structure at different spatial granularities . Coarse-level grids can cancel out high-frequency noise while fine-level grids can capture detailed local patterns, thus enhancing distinctiveness. Third, LDB selects a subset of highly-variant and distinctive bits and concatenates them to form a compact and unique LDB descriptor. .

4.1.13 LATCH

Learned Arrangements of Three Patch Codes descriptors (LATCH) is a fast and compact binary descriptor used to represent local image regions. It describs local image appearances using binary strings since binary descriptors are known for speed and low memory constraints. A known shortcoming of these representations is their inferior performance compared to larger, histogram based descriptors such as the SIFT. Because existing binary descriptors are at an increased risk from noise and local appearance variations, this method helps to optimize the descriptors to handle this conditions.

4.1.14 BinBoost

Boosting Binary Keypoint Descriptors (BinBoost) provides an efficient alternative to other floating-point competitors as it enables faster processing while requiring less memory.It consists of a framework to learn an extremely compact binary descriptor 'BinBoost', that is very robust to illumination and viewpoint changes. Each bit of the descriptor is computed with a boosted binary hash function, and it is shown how to efficiently optimize the different hash functions so that they complement each other, which is key to compactness and robustness. The hash functions rely on weak learners that are applied directly to the image patches which automatically helps to learn the image gradient pooling configuration of the final descriptor. The resulting descriptor significantly outperforms the state-of-the-art binary descriptors and performs similarly to the best floating-point descriptors at a fraction of the matching time and memory footprint.

4.1.15 FREAK

A cascade of binary strings is computed by efficiently comparing image intensities over a retinal sampling pattern called FREAK.

5 Deep Learning Feature Detection and Matching Algorithms

5.1 LF-Net

Local Features Network(LF-Net) proposes a deep learning architecture to learn to detect keypoints and create a descriptor for the detected keypoints. Keypoint matching mainly involves three steps: 1) Keypoint detection, 2) Estimation of scale and orientation, 3) Descriptor creation. This paper proposes an intelligent way of using deep networks to achieve these 3 steps in an end-to-end fashion.

5.2 LIFT

Learned Invariant Feature Transform (LIFT) consists of a Deep Network architecture that implements the full feature point handling pipeline, that is, detection, orientation estimation, and feature description. It consists of a single model that does all three in a unified manner while preserving end-to-end differentiability.

5.3 TILDE

Temporally Invariant Learned DEtector (TILDE) involves a learning-based approach to detect repeatable keypoints under drastic imaging changes of weather and lighting conditions to which state-of-the-art keypoint detectors are surprisingly sensitive.
Good keypoint candidates are first identified in multiple training images taken from the same viewpoint, then a regressor is trained to predict a score map whose maxima are those points so that they can be found by simple non-maximum suppression.

5.4 DeepDesc

Discriminative Learning of Deep Convolutional Feature Point Descriptors (DeepDesc) uses Convolutional Neural Networks (CNNs) to learn discriminant patch representations through training a Siamese network with pairs of (non-)corresponding patches. By using the L2 distance during both training and testing a 128-D descriptors was developed whose euclidean distances reflect patch similarity, and which can be used as a drop-in replacement for any task involving SIFT. The descriptors generalize well against scaling and rotation, perspective transformation, non-rigid deformation, and illumination changes.

5.5 ConvOpt

Learning Local Feature Descriptors Using Convex Optimisation (ConvOpt) learns descriptors suitable for the sparse feature detectors used in viewpoint invariant matching.
First, the learning of the pooling regions for the descriptor is formulated as a convex optimisation problem selecting the regions using sparsity.
Second, the descriptor dimensionality reduction is also formulated as a convex optimisation problem, using Mahalanobis matrix nuclear norm regularisation.
Both formulations are based on discriminative large margin learning constraints.
iIt is extended to a weakly supervised case, which allows to learn the descriptors from unannotated image collections

5.6 DSP-SIFT

Domain-Size Pooling in Local Descriptors(DSP-SIFT) introduces a simple modification of local image descriptors, such as SIFT, based on pooling gradient orientations across different domain sizes, in addition to spatial locations. The resulting descriptor, DSP-SIFT, outperforms other methods in wide-baseline matching benchmarks, including those based on convolutional neural networks, despite having the same dimension of SIFT and requiring no training. It DSP-SIFT, pooling occurs across different domain sizes, Patches of different sizes are re-scaled,gradient orientation computed,pooled across locations and scales and concatenated yielding a descriptor of the same dimension of ordinary SIFT

5.7 SuperPoint

Self-Supervised Interest Point Detection and Description (SuperPoint) is a self-supervised framework for training interest point detectors and descriptors suitable for a large number of multiple-view geometry problems in computer vision. As opposed to patch-based neural networks, it uses a fully-convolutional model operated on full-sized images (VGG-Encoder) and jointly computes pixel-level interest point locations and associated descriptors in one forward pass (Seperate decoders for each). They also introduced Homographic Adaptation, a multi-scale, multi-homography approach for boosting interest point detection repeatability and performing cross-domain adaptation (e.g., synthetic-to-real)

6 Homography

6.1 How are images captured by cameras with same center related?

Images captured by cameras with the same center are related by the Homography equation :

$$\mathbf{H} = \mathbf{K_1^{-1}} \mathbf{R} \mathbf{K_2^{-1}} \tag{1}$$

where K_1 and K_2 are Camera1 and Camera2 Intrinsic Parameters, R is the relative Rotational Matrix

Proof:
Given X and x are homogeneous 3D and 2D points,

For Image 1,

$$x_1 = z^{-1} K_1 [R_1 t_1] X \tag{2}$$

For Image 2,

$$x_2 = z^{-1} K_2 [R_2 t_2] X \tag{3}$$

where R, t are camera extrinsics.

$$x = [K_2][R_2] X \tag{4}$$

As homogeneous, $z = 1$, and an Camera's origin coincide, $t = 0$ Hence, taking inverse of the above equation, and substituting x,

$$
\begin{aligned}
x_1 &= [K_1 R_1] X \\
x_2 &= [K_2 R_2] X \\
\Rightarrow x_2 &= [K_2 R_2][K_1^{-1} R_1^{-1}] x_1 \\
\Rightarrow x_2 &= [K_2 (R_2 / R_1) K_1^{-1}] x_1 \\
\Rightarrow x_2 &= [K_2 R K_1^{-1}] x_1
\end{aligned}
$$

where R is the relative rotational matrix between the two images

Thus, $x_i = HX$ where, $H = [K_2 R K_1^{-1}]$. and H is called the Homography Matrix.

7 References