

UDK 004.93

DOI: 10.37069/1683-4720-2025-39-8

©2025. O. Samoilenko

MATHEMATICAL ANALYSIS OF IMAGE MATCHING TECHNIQUES

Image matching is a fundamental problem in Computer Vision with direct applications in robotics, remote sensing and geospatial data analysis. The present paper reports an analytical and experimental evaluation of classical local feature-based image matching algorithms in the context of satellite imagery, specifically, the Scale-Invariant Feature Transform (SIFT) and the Oriented FAST and Rotated BRIEF (ORB) that is built on the Features from Accelerated Segment Test (FAST) keypoint detector and the Binary Robust Independent Elementary Features (BRIEF) descriptor. Each method is decomposed into a common matching pipeline consisting of the keypoint detection, local descriptor extraction, descriptor matching, and geometric verification. Formally, an image with c color channels is modeled as a function $I : \mathbb{R}^2 \rightarrow \mathbb{R}^c$. Keypoints (features) are locally distinctive points in the image, defined as spatial locations $\mathbf{x} \in \mathbb{R}^2$. Each keypoint is associated with a compact descriptor vector $f(\mathbf{x}) \in \mathbb{R}^D$, where D denotes the descriptor dimensionality. The descriptor matching stage attempts to establish a correspondence $f(\mathbf{x}_i) \approx f(\mathbf{x}'_i)$ between sets of points according to a predefined distance metric, such that the pair $(\mathbf{x}_i, \mathbf{x}'_i)$ represents the same physical location. To ensure geometric consistency, we incorporate a verification employing the Random Sample Consensus (RANSAC) algorithm, where a homography matrix $\mathbf{H} \in \mathbb{R}^{3 \times 3}$ is estimated to model a projective transformation between the images. A match is denoted as an inlier (or valid match) if it satisfies the condition $\|\mathbf{x}'_i - \mathbf{H}\mathbf{x}_i\| < \epsilon$ for a predefined threshold ϵ . The proportion of correspondences satisfying this constraint is referred to as the Inlier Ratio, is used as a measure of matching confidence and serves as the primary metric in the evaluation. The study utilizes a manually constructed dataset of satellite images. The dataset includes GPS-annotated map tiles with intentional partial overlaps between adjacent images enabling a reliable evaluation via a pairwise image matching. We examine the impact of varying the number of extracted keypoints on the resulting Inlier Ratio.

MSC: 68T45, 68U10, 65D19.**Keywords:** image matching, feature detection, computer vision, pattern recognition.

1. Introduction.

Image matching is a fundamental problem in Computer Vision. A particularly challenging and important subproblem involves matching the satellite images due to repetitive patterns (e.g. grid-like street layouts, rows of buildings, or agricultural fields) which can lead to a high number of ambiguous matches, illumination and weather variations, temporal dynamics (e.g. new construction or demolition), limited texture in certain terrains (e.g. deserts, forests, water bodies) where distinctive keypoints may be sparse or nonexistent.

The motivation for the matching map images arises from numerous real-world applications in robotics, remote sensing, and visual geolocatization. Given the visual variability of the map data, robust matching techniques should be designed to handle the affine distortions, noise, and limited texture information. This stimulates the inves-

The authors confirm that there is no conflict of interest and acknowledges financial support by the Simons Foundation grant (SFI-PD-Ukraine-00014586, O.S.) and the National Academy of Sciences of Ukraine under the budget programme 0125U000299.

tigation and evaluation of the feature matching methods, specifically for cartographic imagery. Understanding the strengths and limitations of these methods in the context of the map matching can significantly contribute to improvements in the geospatial data processing and automated map analysis systems.

To evaluate and benchmark the classical image matching techniques, we constructed a custom dataset (Figure 1) consisting of a satellite imagery collected by utilizing the Google Maps Static API [1], capturing images tile-by-tile over a predefined geographic region. Each image is defined by the GPS coordinates of its center point and a fixed zoom level ensuring a consistent spatial resolution, scale and minimal distortion across the dataset. The sampling grid was designed to ensure a partial overlap between the neighbouring images enabling a robust pairwise matching evaluation.



Figure 1. *Satellite Image Dataset. The images are sampled such that the neighbouring images have overlapping regions, enable evaluation of the image matching and alignment algorithms. A collection of the satellite images is retrieved by utilizing the Google Maps Static API [1].*

The present work aims to provide a rigorous analysis of the local image matching techniques with a particular focus on classical local feature-based descriptors such as the Scale-Invariant Feature Transform (SIFT) [2] and the Oriented FAST and Rotated BRIEF (ORB) [3] that is built on the Features from Accelerated Segment Test (FAST) keypoint detector [4] and the Binary Robust Independent Elementary Features (BRIEF) descriptor [5]. We also consider their integration with matching algorithms and geometric verification pipelines, such as the Random Sample Consensus (RANSAC) [6], as well as deliver a qualitative evaluation of their accuracy in the satellite imagery domain.

A number of studies have explored the descriptors evaluation in various contexts. Mikolajczyk et al. [7] compared local descriptors under image transformations. Panchal et al. [8] analysed the SIFT and other algorithms based on their performance in the keypoint detection and descriptor matching. Their evaluation conducted on a single dataset without cross-view variations, focused on the computation time, the number of detected keypoints, and the matching accuracy under affine and scale transformations. Whereas the study provides valuable insights into the strengths and weaknesses of these classical methods in a controlled setting, it does not address more complex scenarios such as cross-view which are crucial for geo-localization and place recognition tasks.

Benchmarks such as in [9] provide the standardized evaluation frameworks. Recent comparative studies [10] have extensively benchmarked handcrafted descriptors on standard datasets highlighting their trade-offs in terms of the accuracy and computational cost.

The main contributions of the present paper are as follows:

- We formulate the mathematical foundations underlying classical handcrafted keypoint descriptors including the SIFT and ORB.
- We evaluate the accuracy for image matching pipelines based on the RANSAC inliers including the Inlier Ratio metric.
- We conduct a systematic numerical comparison of local descriptors analyzing their performance under different accuracy metrics.

The paper structure is organized in the following way: Section 2 introduces mathematical formulation of the digital image and notations adopted throughout the text. Section 3 defines the concept of the keypoint and local descriptor formulation, provides an overview of the SIFT and ORB methods emphasizing their properties and differences. Section 4 describes the keypoint matching procedure based on the descriptor similarity and presents the RANSAC algorithm for the geometric outlier rejection. Section 5 outlines the accuracy metrics used to evaluate the keypoint matching quality. Section 6 presents a numerical evaluation of the discussed methods on a benchmark dataset with analysis of the matching performance. In Section 7, the summary is provided and future research directions are outlined.

2. Image Definition.

The image can be mathematically defined as a function,

$$I : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^c,$$

where the continuous domain Ω represents an image in the OXY-plane but the range $\mathbb{R}^c$ represents an intensity value with c channels (e.g. $c = 1$ for the grayscale image, $c = 3$ for the RGB image). In the discrete case, the digital image is modeled as

$$I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathbb{R}^c,$$

where the discrete domain $\mathcal{D} = \{0, \dots, H - 1\} \times \{0, \dots, W - 1\}$ is the pixel locations grid, $W > 0$ and $H > 0$ are the width and height of the image in pixels, respectively, the origin $O = (0, 0)$ is the top-left pixel, $I(x, y) \in \mathbb{R}^c$ returns the pixel value (intensity or color vector) at the integer location (x, y) .

3. Scale-Invariant Feature Transform (SIFT).

The keypoint (also called the interest point or feature point) is a distinctive and repeatable location in an image that can be used for matching points between images. Desirable properties of the ideal keypoint are repeatability (the same keypoint should be found in multiple views of the same scene). It should have a unique distinctive local

neighborhood and be invariant to scale changes, rotation, illumination and viewpoint transformations.

Early primitive keypoint detectors were focusing on finding locally distinctive points like corners or blobs without a full invariance to the scale or rotation. One of the earliest and most influential methods is the Harris corner detector [11], which identifies points where the image intensity strongly changes in two directions – typically at corners. The detection exploits the Moravec operator [12], which detects interest points based on shifting the windows in multiple directions. Another early approach is the Laplacian of Gaussian [13], which detects blob-like structures by locating the extrema in a scale-normalized Laplacian. These detectors provide repeatable and distinctive keypoints but lack the scale or rotation invariance – making them less robust under transformations compared to later methods like the SIFT.

The SIFT descriptor encodes local appearance of an image region surrounding a keypoint in a manner that is invariant to the scale, rotation, and moderate affine distortions. The descriptor is constructed through the following steps:

Scale-Space Construction: In the SIFT, the Gaussian function with the kernel

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

is adopted, where the standard deviation σ controls amount of blurring and plays central role in keypoints detection. The Gaussian blur smooths the image and reduces the high-frequency noise, thereby improving the stability of the keypoint detection.

A scale-space representation is created by convolving the image with the Gaussian kernels at different scales resulting in multi-scale image pyramid. The continuous convolution in the scale-space construction is defined by

$$L(x, y, \sigma) = \iint_{\mathbb{R}^2} I(\xi, \eta) G(x - \xi, y - \eta, \sigma) d\xi d\eta,$$

where the variables ξ and η represent coordinates in a neighborhood of the point (x, y) . They are used to compute the weighted sum of the image values by the Gaussian function. Pixels closer to the center (x, y) contribute more due to the higher Gaussian values. In the continuous space, the domain of integration is $\mathbb{R}^2$, i.e., the entire image plane. In discrete implementations, the convolution with the kernel size k takes the form

$$L(x, y, \sigma) = \sum_{i=-k/2}^{k/2} \sum_{j=-k/2}^{k/2} I(x - i, y - j) G(i, j, \sigma)$$

or, simply,

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y),$$

where $*$ is the convolution operator.

Keypoint Detection: Keypoints are identified as local extrema in the Difference of Gaussians function by repeatedly comparing each point with its neighbors across both spatial and scale dimensions, increasing the standard deviation σ producing different levels of blur (Figure 2):

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma),$$

where k is a multiplicative factor, typically $\sqrt{2}$.

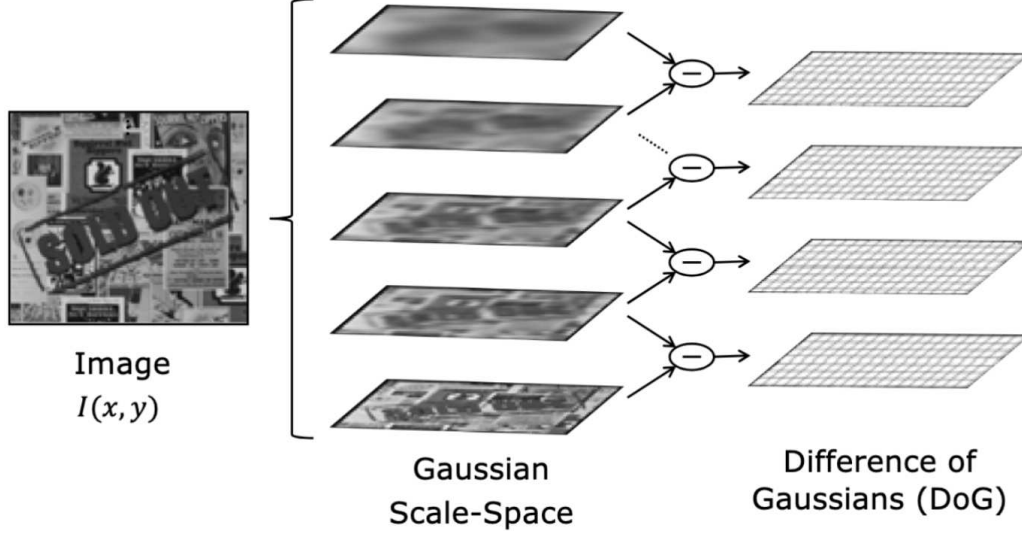


Figure 2. Difference of the Gaussians [14] is computed by subtracting the scale-space representation of an input image at different levels of the Gaussian blurring resulting in a pyramid of images, which is utilized for detecting the scale-invariant keypoints.

In the SIFT, a keypoint is considered if a pixel in the Difference of Gaussians image pyramid is a local extremum with respect to its neighborhood in the scale and space, that is, its value is either greater or less than all the surrounding pixels in both spatial and scale dimensions.

Orientation Assignment: The orientation assignment step in the SIFT is crucial to achieve rotation invariance of keypoints. To make descriptors invariant to the image rotation, each keypoint is assigned to a dominant orientation based on the local image gradients. The descriptor is then computed relative to this orientation so that the image rotation does not affect the descriptor. This means, no matter how the image is rotated, the descriptor will look in the same way.

For each keypoint, an orientation histogram is constructed from the gradient magnitudes and orientations within a local neighborhood. The gradient magnitude and direction at each pixel are computed by

$$m(x, y) = \sqrt{L_x^2 + L_y^2} \quad \text{and} \quad \theta(x, y) = \arctan\left(\frac{L_y}{L_x}\right),$$

respectively, where L_x and L_y denote the image derivatives in the x and y directions, respectively, computed from the smoothed image $L(x, y, \sigma)$. The dominant orientation is selected as the histogram peak keeping the rotation invariance.

Descriptor Vector Construction: In the SIFT descriptor construction, a local image patch, which is centered at a detected keypoint, is extracted and processed to capture distinctive gradient-based information. Around each detected keypoint (x_p, y_p) , a square patch Ω_p of the size $S \times S$ is considered. The patch Ω_p is divided into the $M \times M$ spatial subregions of the size $s \times s$, where $s = S/M$. In each subregion, an orientation histogram is constructed. Denote the bin number as B covering the angular range $[0, 2\pi)$ with the bin centers

$$\theta_b = \frac{2\pi b}{B}, \quad b = 0, 1, \dots, B-1.$$

Each pixel (x, y) within a patch contributes to the histogram with a vote weighted by its gradient magnitude and the spatial Gaussian weighting function

$$w(x, y) = \exp\left(-\frac{(x - x_p)^2 + (y - y_p)^2}{2\sigma^2}\right).$$

The orientation histogram for the subregion r is as follows:

$$H_b^r = \sum_{(x,y) \in r} w(x, y) \cdot m(x, y) \cdot \mathbf{1}_b(\theta(x, y)), \quad b = 1, 2, \dots, B,$$

where $m(x, y)$ is the gradient magnitude, $\mathbf{1}_b(\theta)$ is an indicator of assigning the gradient orientation to the bin b ,

$$\mathbf{1}_b(\theta) = \begin{cases} 1, & \theta \in \left[\frac{2\pi(b-1)}{B}, \frac{2\pi b}{B}\right) \\ 0, & \text{otherwise.} \end{cases}$$

After computing the histograms for all $M \times M$ subregions, they are concatenated into the single descriptor vector

$$\mathbf{d}_{\text{SIFT}} = [H_0^1, \dots, H_{B-1}^1, H_0^2, \dots, H_{B-1}^{M^2}]. \quad (1)$$

Thus, the total dimensionality of the SIFT descriptor is $D = M^2 \times B$. In practice, the SIFT descriptor is commonly constructed using a local image patch of the size 16×16 pixels centered at the keypoint. This patch is divided into a grid of 4×4 spatial subregions, and each subregion contributes a histogram of gradient orientations. Each orientation histogram typically contains 8 bins. As a result, the final descriptor is a concatenation of 4×4 histograms, each with 8 bins yielding a feature vector of the dimension $D = 4^2 \times 8 = 128$.

4. Oriented FAST and Rotated BRIEF (ORB).

The ORB descriptor is a fast and efficient alternative to the SIFT combining a keypoint detector based on the FAST algorithm with an orientation component and a binary descriptor derived from BRIEF. It is particularly suitable for real-time applications and devices with limited computational resources.

Keypoint Detection using the FAST: The ORB employs the FAST detector, which performs binary intensity tests around a circular region. A point (x, y) is a FAST keypoint if there exists a contiguous arc of, at least, N pixels in the Bresenham circle centered at (x, y) (Figure 3) that are either significantly brighter or darker than the center point, i.e.,

$$I(x_q, y_q) > I(x, y) + t \quad \text{or} \quad I(x_q, y_q) < I(x, y) - t, \quad q = 1, 2, \dots, B,$$

where t is an intensity threshold, $(x_q, y_q)_{q=1}^B$ is the set of B points lying on the Bresenham circle centered at (x, y) . The use of a circular patch is important because it provides rotational symmetry around the candidate pixel ensuring that the detection is invariant to image rotations. This results in the high-speed feature detection but without a strength score.

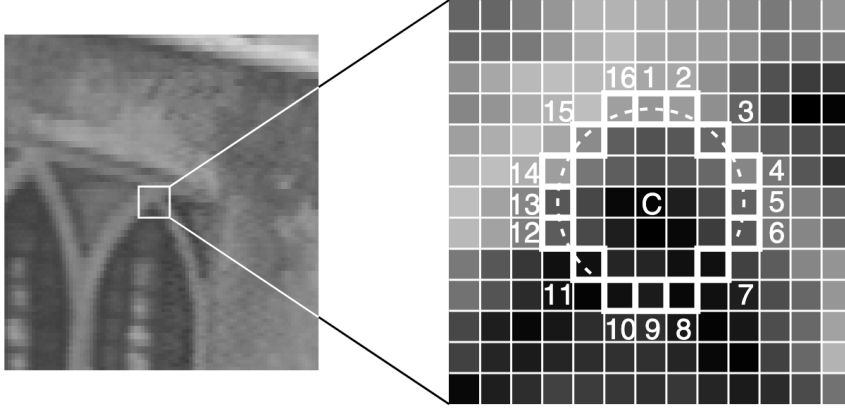


Figure 3. The FAST feature detection [4]. As example, the Bresenham circle of the radius 3 with the center at C is presented. The highlighted squares are the pixels adopted in the feature detection.

To rank and filter the FAST keypoints, the ORB computes the Harris corner measure [11] by using the second moment matrix

$$M = \begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix}, H(x, y) = \det(M) - \alpha \cdot (\text{trace}(M))^2,$$

where I_x and I_y denote the image derivatives in horizontal and vertical directions respectfully; α is a sensitivity constant, typically, $\alpha \in [0.04, 0.06]$.

Rotated BRIEF Descriptor: The ORB adopts the BRIEF descriptor, which relies on intensity comparisons between the point pairs. Typically, the pairs are drawn randomly according to an Gaussian distribution centered at the keypoint with zero mean and standard deviation proportional to the patch size. This strategy ensures that the pairs are well distributed spatially and do not cluster only near the center. Let a local patch Ω centered at a keypoint be defined on a square domain of the size $S \times S$ pixels. In the BRIEF descriptor, we sample N pairs of the pixel coordinates

$$\{(\mathbf{p}_i, \mathbf{q}_i) \mid \mathbf{p}_i, \mathbf{q}_i \in \Omega \subset \mathbb{R}^2\}_{i=1}^N.$$

Each pair defines a binary intensity test,

$$\tau_i = \begin{cases} 1, & I(\mathbf{p}_i) < I(\mathbf{q}_i) \\ 0, & \text{otherwise.} \end{cases}$$

Repeating this test for N pairs yields the N -bit binary BRIEF descriptor

$$\mathbf{d}_{\text{BRIEF}} = [\tau_1, \tau_2, \dots, \tau_N] \in \{0, 1\}^N. \quad (2)$$

To achieve the rotation invariance, the test pairs are rotated by the dominant orientation angle $\theta \in [0, 2\pi)$ which is selected by computing a histogram of gradient orientations within the image patch around the keypoint

$$\mathbf{p}'_i = \mathbf{R}_\theta \mathbf{p}_i, \quad \mathbf{q}'_i = \mathbf{R}_\theta \mathbf{q}_i, \quad \mathbf{R}_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix},$$

where $\mathbf{R}_\theta$ is the rotation matrix, $(\mathbf{p}'_i, \mathbf{q}'_i)$ are the resulting rotated coordinates to ensure the rotation invariance.

5. Keypoint Matching Techniques.

Image matching is the task of establishing visual correspondences between two images by comparing local features. The standard pipeline consists of detecting the keypoints, extracting the local descriptors, matching these descriptors, and verifying them by using geometric constraints. To compare the descriptors extracted from images, we employ standard similarity measures.

SIFT Descriptor Matching: Commonly used metric for the SIFT is the Euclidean distance (also called L_2 -norm). Euclidean distance measures the straight-line distance between two descriptors. Given two descriptors $\mathbf{d}_1, \mathbf{d}_2 \in \mathbb{R}^D$, the Euclidean distance is defined as

$$d_{\text{Euclidean}}(\mathbf{d}_1, \mathbf{d}_2) = \|\mathbf{d}_1 - \mathbf{d}_2\|_2 = \sqrt{\sum_{i=1}^D (\mathbf{d}_{1,i} - \mathbf{d}_{2,i})^2}, \quad (3)$$

where $\mathbf{d}$ is the SIFT descriptor defined in (1) and D is the descriptor dimensionality.

BRIEF Descriptor Matching: The BRIEF descriptors are matched by utilizing the Hamming distance. The Hamming distance is defined as the number of positions at which their bits differ, i.e.,

$$d_{\text{Hamming}}(\mathbf{d}_1, \mathbf{d}_2) = \sum_{i=1}^N [\mathbf{d}_{1,i} \neq \mathbf{d}_{2,i}], \quad (4)$$

where $\mathbf{d}$ is the BRIEF descriptor defined in (2), N is the descriptor dimensionality and $[\cdot]$ indicates whether the i -th bits of the binary descriptor vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ differ. In other words, the Hamming distance is simply the count of the bit mismatches between the two binary strings, making it computationally efficient for comparing the BRIEF descriptors.

Descriptor Matching with Brute Force: Let $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$ and $\{y_j\}_{j=1}^M \subset \mathbb{R}^D$ be the sets of descriptors extracted from the first and second image, respectively, where D is the descriptor dimension. The Brute Force Matcher compares each descriptor x_i with every y_j and finds the best match j^* for each i by minimizing a distance metric d :

$$j^* = \arg \min_j d(\mathbf{d}_i, \mathbf{d}_j).$$

where distance metric is $d_{\text{Euclidean}}$ (3) for the SIFT and d_{Hamming} (4) for the ORB. We define the set of all the Brute Force matches as

$$\mathcal{M} = \left\{ (\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_j = \arg \min_j d(\mathbf{d}_i, \mathbf{d}_j) \right\}. \quad (5)$$

Many incorrect matches (outliers) are present due to the lack of geometric verification (Figure 4).

Homography Estimation: A homography is a projective transformation that maps points from one image plane to another under perspective geometry. It is represented by a 3×3 matrix $\mathbf{H} \in \mathbb{R}^{3 \times 3}$, and is defined up to a non-zero scale factor. Given a point $\mathbf{x} = (x, y, 1)$ in the homogeneous coordinates, the homography relates it to its image $\mathbf{x}' = (x', y', 1)$ in the second view as

$$\lambda \mathbf{x}' = \mathbf{H} \mathbf{x}$$

where λ is a nonzero scale factor (due to homogeneous coordinates) and

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

is the parameterized homography with elements h_{ij} that should be found from a set of corresponding points between two images.

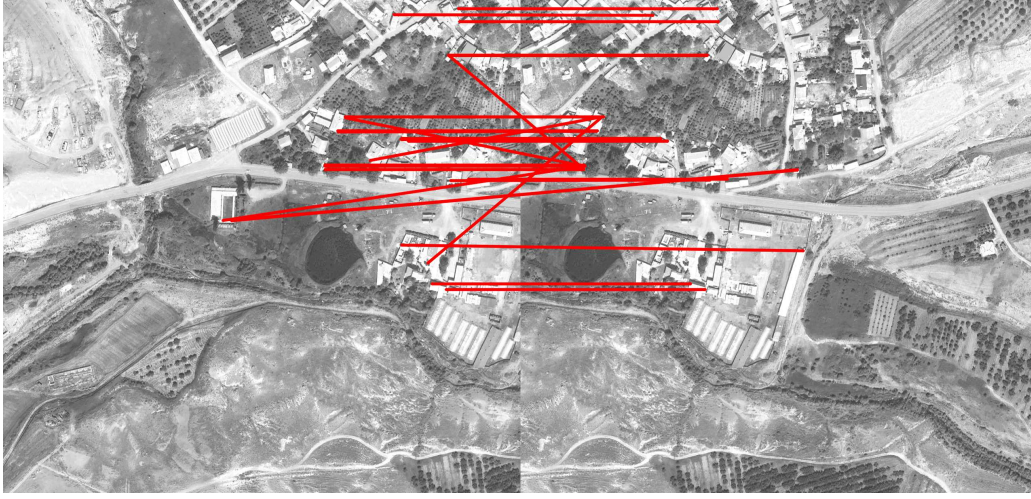


Figure 4. Keypoint matches with the Brute Force. Initial keypoint correspondences between two images using a Brute Force Matcher on SIFT descriptors, including outliers. Images are retrieved by utilizing the Google Maps Static API [1].

For each point correspondence $(x, y) \leftrightarrow (x', y')$, the projective mapping yields

$$\begin{aligned} x'(h_{31}x + h_{32}y + h_{33}) &= h_{11}x + h_{12}y + h_{13}, \\ y'(h_{31}x + h_{32}y + h_{33}) &= h_{21}x + h_{22}y + h_{23}. \end{aligned}$$

Rewriting the above as linear equations in the unknown vector $\mathbf{h} \in \mathbb{R}^9$, where

$$\mathbf{h} = [h_{11} \ h_{12} \ h_{13} \ h_{21} \ h_{22} \ h_{23} \ h_{31} \ h_{32} \ h_{33}]^\top,$$

we obtain the following linear system for each correspondence:

$$\begin{bmatrix} x & y & 1 & 0 & 0 & 0 & -xx' & -yy' & -x' \\ 0 & 0 & 0 & x & y & 1 & -xy' & -yy' & -y' \end{bmatrix} \cdot \mathbf{h} = \mathbf{0}.$$

The matrix $\mathbf{H}$ has 8 degrees of freedom. It is usually normalized by assuming $h_{33} = 1$. Stacking these equations for $n \geq 4$ point correspondences yields the linear system

$$A\mathbf{h} = \mathbf{0}, \quad A \in \mathbb{R}^{2n \times 9}.$$

Solving the system provides the elements h_{ij} , which are then reshaped into the homography matrix $\mathbf{H}$.

Geometric Verification with the RANSAC: In practice, point correspondences often include outliers due to mismatches. To ensure spatial consistency, the matched keypoints are verified by estimating a geometric transformation. To estimate $\mathbf{H}$ robustly in the presence of outliers, the RANSAC algorithm is adopted. It randomly samples

subsets of matches, computes candidate homographies, and selects the one that maximizes the number of inliers – matches that satisfy

$$\|\mathbf{x}'_i - \mathbf{H}\mathbf{x}_i\| < \epsilon$$

for a predefined threshold ϵ . Repeat the above steps for a fixed number of iterations or until a satisfactory consensus is found. This approach effectively minimizes the influence of outliers and improves the matching accuracy.

Final Inlier Matches: The final set of matches taken for the image alignment or stitching includes only those that are consistent with the estimated homography,

$$\mathcal{M}_{\text{Inlier}} = \{(\mathbf{x}_i, \mathbf{x}'_i) \mid \|\mathbf{x}'_i - \mathbf{H}\mathbf{x}_i\| < \epsilon\}, \quad (6)$$

where ϵ is the reprojection error threshold. It specifies the maximum allowed deviation between the projected point $\mathbf{H}\mathbf{x}_i$ and the observed point $\mathbf{x}'_i$ for the correspondence to be classified as an inlier. Smaller values of ϵ yield stricter geometric consistency at the cost of rejecting more matches. This geometric verification step significantly reduces false matches and improves the reliability of downstream tasks such as image registration and structure-from-motion (Figure 5).

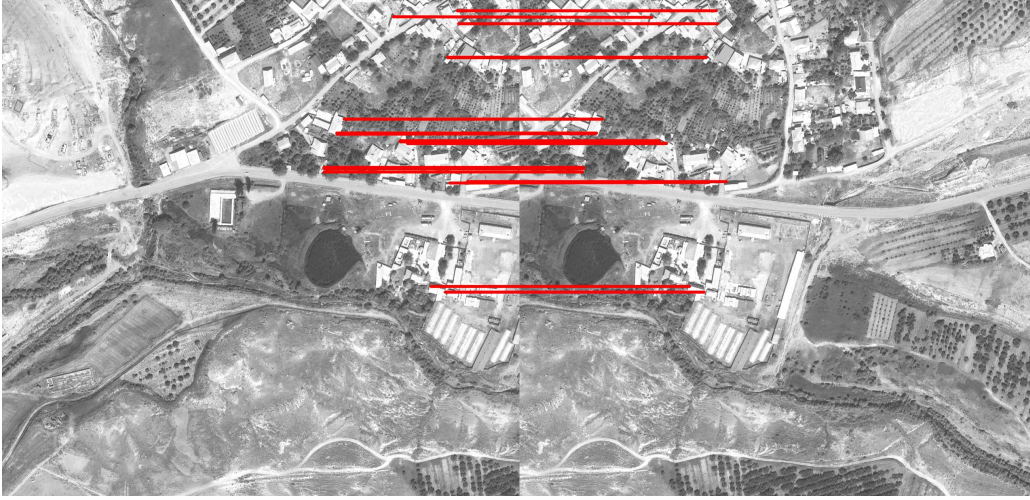


Figure 5. The keypoint matches between two images after applying the RANSAC filtering. The RANSAC rejects outliers and retains only geometrically consistent matches (inliers), leading to a much cleaner and structurally meaningful correspondence set. Images are retrieved by utilizing the Google Maps Static API [1].

6. Accuracy Metrics.

In the image matching pipelines based on local descriptors and geometric verification, one common approach to assess correctness is through the analysis of RANSAC inliers – matches that conform to a geometric model under a reprojection error threshold.

Geometric Inlier Definition: Let $\mathcal{M}_{\text{Inlier}}(I_i, I_j) \subset \mathcal{M}(I_i, I_j)$ denote the set of inliers between images I_i and I_j , where $\mathcal{M}$ is defined in (5) and $\mathcal{M}_{\text{Inlier}}$ is defined in (6). The following metrics can be used to quantify the match quality:

Inlier Ratio: Computes the fraction of total matches that are inliers,

$$R_{\text{Inlier}} = \frac{|\mathcal{M}_{\text{Inlier}}|}{|\mathcal{M}|},$$

where $\mathcal{M}$ and $\mathcal{M}_{\text{Inlier}}$ are defined in (5) and (6), respectively.

Matching Decision Rule: Define a Prediction function $\mathcal{P}$ with threshold ρ computed for the given pair of images by using the matching pipeline,

$$\mathcal{P}_\rho(I_i, I_j) = \begin{cases} 1, & \text{if } R_{\text{Inlier}}(I_i, I_j) \geq \rho \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where $\rho \in (0, 1)$ is a the minimum Inlier Ratio. An image pair is considered a correct match if $\mathcal{P}_\rho(I_i, I_j) = 1$.

Ground truth: The evaluation of the feature matching performance relies on comparing the predicted matches to the true relationships between images. The ground truth refers to the known, correct match status of each image pair in the dataset. In the context of visual place recognition on the map images, the ground truth defines whether two images represent the same location (match) or different locations (non-match). These labels are typically obtained from the manual annotation, spatial metadata, or reliable prior mapping,

$$\mathcal{G}(I_i, I_j) = \begin{cases} 1, & I_i \text{ and } I_j \text{ represent the same location,} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Based on the Ground Truth and Prediction functions, each matched image pair is categorized in the following way:

$$\begin{aligned} \text{True Positive: TP} &= \sum_{i,j} [\mathcal{G}(I_i, I_j) = 1 \wedge \mathcal{P}_\rho(I_i, I_j) = 1], \\ \text{True Negative: TN} &= \sum_{i,j} [\mathcal{G}(I_i, I_j) = 0 \wedge \mathcal{P}_\rho(I_i, I_j) = 0], \\ \text{False Positive: FP} &= \sum_{i,j} [\mathcal{G}(I_i, I_j) = 0 \wedge \mathcal{P}_\rho(I_i, I_j) = 1], \\ \text{False Negative: FN} &= \sum_{i,j} [\mathcal{G}(I_i, I_j) = 1 \wedge \mathcal{P}_\rho(I_i, I_j) = 0], \end{aligned}$$

where $\mathcal{P}$ and $\mathcal{G}$ are defined in (7) and (8), respectively and $[\cdot]$ indicates if the logical expression in brackets is true.

7. Comparative Study.

To evaluate the performance of local feature descriptors for the image matching, we conducted a systematic comparison of the SIFT and ORB based on the Inlier Ratio. This metric provides a direct and interpretable measure of the matching reliability, especially in the absence of the ground truth keypoint correspondences. Each image is compared one-by-one against the entire set of reference images. The ground truth (whether the two map images truly match the same place) provides the correct label. A feature-based matching pipeline is applied consisting of the keypoint detection, descriptor extraction, descriptor matching, and geometric verification using the RANSAC. For each image pair, we estimate homography and count the number of inlier correspondences as a measure of matching confidence. We focus on Inlier Ratio alone and do not compute standard classification metrics such as precision, recall, or F_1 -score. This is because the True Positives naturally exhibit high Inlier Ratios, while the True Negatives exhibit near-zero Inlier Ratios, making binary classification thresholds unnecessary. Experiments were conducted by varying the number of keypoints extracted per image: 100, 200, 500, 1000, and 2000. For each configuration, the mean Inlier Ratio was computed across all image pairs in the dataset. Both ORB and SIFT were tested under identical conditions (Table 1).

Algorithm	Inlier Ratio (True Positives)	Inlier Ratio (True Negatives)
ORB (100)	24.94%	0.52%
ORB (200)	26.48%	0.61%
ORB (500)	27.77%	0.41%
ORB (1000)	28.54%	0.23%
ORB (2000)	29.34%	0.14%
SIFT (100)	32.21%	0.58%
SIFT (200)	32.81%	0.63%
SIFT (500)	34.45%	0.43%
SIFT (1000)	35.29%	0.26%
SIFT (2000)	36.13%	0.15%

Table 1. Comparison of the SIFT and ORB methods based on the matching accuracy measured by inlier correspondences, in %. The number next to each method indicates the number of keypoints extracted per image.

The experimental results demonstrate that the SIFT consistently outperforms the ORB across all feature counts in terms of the Inlier Ratio. Even when extracting only 100 keypoints, the SIFT achieves higher matching accuracy than the ORB at 2000 keypoints, emphasizing its robustness and suitability for the satellite image alignment tasks. Moreover, the performance gains of increasing the features number are not significant after a certain threshold. For instance, when using the SIFT, increasing the features number from 1000 to 2000 results in a noticeable improvement in Inlier

Ratio from 35.29% to 36.13% – less than 1% gain, despite doubling the number of extracted keypoints. The experiments suggest that extracting 200 to 500 keypoints strikes an effective trade-off between the accuracy and computational cost, making these configurations particularly attractive for deployment on resource-constrained platforms, while extracting more features may not be justified.

8. Conclusions.

A comparative analysis of classical feature matching algorithms applied to the satellite imagery is presented. A particular focus was placed on evaluation of the matching accuracy using the standard Computer Vision pipelines involving the keypoint detection, descriptor extraction, feature matching, and geometric verification through the RANSAC. Quantitative evaluation was conducted by using the Inlier Ratio metric constructed from ground-truth correspondences between image pairs. We provide a systematic framework for evaluating classical matching techniques. The presented methods and findings can serve as a foundation for further studies of scenarios, where computational and memory resources are limited.

In the future work, we aim to extend this analysis by

- evaluating the cross-view matching between aerial and satellite imagery where scale, and lighting vary significantly,
- incorporating the learning-based methods such as the SuperPoint [15] for improved robustness,
- experimenting with descriptor aggregation techniques such as the VLAD [16] for efficient image retrieval.

Such extensions will further bridge the gap between classical vision techniques and modern learning-based systems in the context of the geospatial image analysis and localization.

References

1. Google. (2025). Maps Static API. *Google Developers*. Retrieved April 10, 2025, from <https://developers.google.com/maps/documentation/maps-static/start>
2. Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60, 91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
3. Rublee, E., Rabaud, V., Konolige, K., Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *Proc. IEEE ICCV*, 2564–2571. <https://doi.org/10.1109/ICCV.2011.6126544>
4. Rosten, E., Drummond, T. (2005). Fusing points and lines for high performance tracking. *Proc. IEEE ICCV*, 2, 1508–1515. <https://doi.org/10.1109/ICCV.2005.104>
5. Calonder, M., Lepetit, V., Strecha, C., Fua, P. (2010). BRIEF: Binary robust independent elementary features. *Proc. European Conf. Computer Vision*, 778–792. https://doi.org/10.1007/978-3-642-15561-1_56
6. Fischler, M.A., Bolles, R.C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395. <https://doi.org/10.1145/358669.358692>
7. Mikolajczyk, K., Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10), 1615–1630. <https://doi.org/10.1109/TPAMI.2005.188>

8. Panchal, P.M., Panchal, S.R., Shah, S.K. (2013). A comparison of SIFT and SURF. *Int. J. Innovative Research in Computer and Communication Engineering*, 1(2), 323–327.
9. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K. (2017). HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. *Proc. IEEE CVPR*, 5173–5182. <https://doi.org/10.1109/CVPR.2017.410>
10. Tareen, S.A.K., Saleem, Z. (2018). A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. *Proc. IEEE iCoMET*, 1–10. <https://doi.org/10.1109/ICOMET.2018.8346440>
11. Harris, C., Stephens, M. (1988). A combined corner and edge detector. *Proc. Alvey Vision Conf.*, 15(50), 10–5244. <https://doi.org/10.5244/C.2.23>
12. Moravec, H.P. (1980). Obstacle avoidance and navigation in the real world by a seeing robot rover. *Stanford University*.
13. Lindeberg, T. (1998). Feature detection with automatic scale selection. *Int. J. Computer Vision*, 30(2), 79–116. <https://doi.org/10.1023/A:1008045108935>
14. Nayar, S.K. (2022). SIFT Detector. *Columbia University*. Retrieved September 15, 2024, from <https://cave.cs.columbia.edu/Statics/monographs/SIFT%20Detector%20FPCV-2-3.pdf>
15. DeTone, D., Malisiewicz, T., Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. *Proc. IEEE CVPR*, 224–236. <https://doi.org/10.1109/CVPRW.2018.00060>
16. Jégou, H., Douze, M., Schmid, C., Pérez, P. (2010). Aggregating local descriptors into a compact image representation. *Proc. IEEE CVPR*, 3304–3311. <https://doi.org/10.1109/CVPR.2010.5540039>

O. Самойленко

Математичний аналіз методів зіставлення зображень.

Зіставлення зображень є фундаментальною проблемою комп'ютерного зору з прямими застосуваннями в робототехніці, дистанційному зондуванні та аналізі геопросторових даних. У цій роботі представлено аналітичну та експериментальну оцінку класичних алгоритмів зіставлення зображень на основі локальних ознак, у контексті супутникових знімків, зокрема Scale-Invariant Feature Transform (SIFT) та Oriented FAST and Rotated BRIEF (ORB), який базується на детекторі ключових точок Features from Accelerated Segment Test (FAST) та дескрипторі Binary Robust Independent Elementary Features (BRIEF). Кожен метод розкладено на послідовність етапів зіставлення, що складається з виявлення ключових точок, обчислення локальних дескрипторів, зіставлення дескрипторів та геометричної перевірки. Формально, зображення з c кольоровими каналами моделюється як функція $I : \mathbb{R}^2 \rightarrow \mathbb{R}^c$. Ключові точки (ознаки) — це локально характерні точки зображення, визначені як координати у просторі $\mathbf{x} \in \mathbb{R}^2$. Кожна ключова точка пов'язана з компактним вектором-дескриптором $f(\mathbf{x}) \in \mathbb{R}^D$, де D позначає розмірність дескриптора. На етапі зіставлення дескрипторів виконується встановлення відповідності $f(\mathbf{x}_i) \approx f(\mathbf{x}'_i)$ між множинами точок згідно з попередньо заданою метрикою відстані, такою, що пара $(\mathbf{x}_i, \mathbf{x}'_i)$ представляє ту саму фізичну точку простору. Для забезпечення геометричної узгодженості ми застосовуємо перевірку за допомогою алгоритму Random Sample Consensus (RANSAC), в якому оцінюється матриця гомографії $\mathbf{H} \in \mathbb{R}^{3 \times 3}$, що моделює проєктивне перетворення між зображеннями. Відповідність вважається валідною, якщо виконується умова $\|\mathbf{x}'_i - \mathbf{H}\mathbf{x}_i\| < \epsilon$ для певного порогового значення ϵ . Частка відповідностей, що задовольняють цю умову, називається коефіцієнтом валідних відповідностей (Inlier Ratio) і використовується як міра впевненості у зіставленні та є основною метрикою оцінювання. У дослідженні використовується вручну сформована вибірка супутникових зображень. Набір даних включає фрагменти карти з прив'язкою до GPS

та навмисним частковим перетином між сусідніми зображеннями, що забезпечує можливість надійної оцінки шляхом попарного зіставлення. Ми досліджуємо вплив кількості обрахованих ключових точок на значення коефіцієнту валідних відповідностей.

Ключові слова: зіставлення зображень, виявлення ознак, комп'ютерний зір, розпізнавання образів.

*Institute of Mathematics, National Academy of Sciences of Ukraine,
Kyiv, Ukraine
oleh.samoilenko@imath.kiev.ua*

Received 14.04.25