UDC 004.93
DOI: https://doi.org/10.17721/1812-5409.2025/2.21

Oleh SAMOILENKO, PhD Student
ORCID ID: 0000-0003-4466-4004
e-mail: oleh.samoilenko@imath.kiev.ua
Institute of Mathematics of the NAS of Ukraine, Kyiv, Ukraine

MATHEMATICAL PROBLEMS OF FEATURE MATCHING FOR VISION-GUIDED VEHICLES WITH LIMITED RESOURCES

Vision-based sensing plays a critical role for autonomous vehicles, where image data serve as the primary input for navigation, mapping, state estimation, and motion control. These operate under real-time and resource-constrained conditions, requiring feature detection and matching algorithms that are accurate and computationally efficient. A key component of such pipelines is the identification of keypoints – distinctive image locations $\mathbf{x} \in \mathbb{R}^2$ ***within a digital image function*** $I : \mathbb{R}^2 \to \mathbb{R}^c$***, where*** c ***is the channels number, and*** $I(\mathbf{x}) \in \mathbb{R}^c$ ***denotes the local pixel intensity. Each detected keypoint is associated with a descriptor*** $\mathbf{d} \in \mathbb{R}^D$ ***encoding a local visual structure in a form invariant to translation, rotation, and moderate scale changes. Keypoints between two images*** $\mathbf{x}_i$ ***and*** $\mathbf{x}'_i$ ***are matched by comparing descriptors yielding the correspondences*** $\mathbf{x}'_i \leftrightarrow \mathbf{x}_i$***. Geometric verification is performed by estimating a transformation matrix*** $\mathbf{H} \in \mathbb{R}^{3\times3}$ ***and accepting*** $\|\mathbf{x}'_i - \mathbf{H}\mathbf{x}_i\| < \varepsilon$***, where*** ε ***is a matching tolerance. The proportion of such inlier correspondences referred to as the inlier ratio serves as an accuracy metric. Classical keypoint methods, e.g. the Scale-Invariant Feature Transform (SIFT) versus learned methods, e.g. the SuperPoint (applied to a manually constructed dataset of satellite imagery), are mathematically evaluated. Performance is analyzed in terms of inlier ratio and computational efficiency, reflecting the trade-offs between robustness and resource use. The results pursue design guidelines for integrating lightweight yet accurate vision algorithms into platforms such as UAVs, enabling reliable visual odometry and the Simultaneous Localization and Mapping (SLAM) under constrained hardware conditions.***

K e y w o r d s: *unmanned autonomous vehicles, image matching, feature detection, computer vision, pattern recognition.*

AMS 2020 classification: 68T45, 68U10, 65D19.

Introduction

Vision-based sensing has become a cornerstone of modern autonomous vehicles to perceive and navigate complex environments, where cameras provide rich environmental information for navigation, mapping, and control, with particular importance in the GPS-denied environments (Boyun et al., 2025). Then the feature detection methods can be divided into two categories: classical and learned ones. Classical techniques, e.g. the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), rely on hand-crafted heuristics such as image gradients and binary intensity tests. They are lightweight and facilitate real-time Simultaneous Localization and Mapping (SLAM) systems (Mur-Artal, Montiel, & Tardos, 2015) and visual odometry computation pipelines (Scaramuzza, & Fraundorfer, 2011). These are efficient on CPU-based embedded systems but could be sensitive to noise, illumination, and geometric distortions. The learned methods, e.g. the SuperPoint (DeTone, Malisiewicz, & Rabinovich, 2018), employ deep neural networks to jointly detect and describe keypoints, as well as learning robust features, directly from the data. Trained on diverse image collections, they achieve higher repeatability and robustness to viewpoint and illumination changes compared to classical handcrafted descriptors, and became widely adopted in remote sensing applications (Rusyn, Lutsyk, & Kosarevych, 2021). Combining with learned matchers like SuperGlue (Sarlin et al., 2020) yields the state-of-the-art results in image matching and relative pose estimation, consistently outperforming the SIFT. Large-scale benchmarks (Huang et al., 2024) consistently report improved reconstruction quality and localization accuracy for learned pipelines. Nevertheless, these improvements come at the expense of computational complexity. The SuperPoint involves a dense convolutional architecture with multiple encoder/decoder stages, requiring on the order of several gigaflops (GFLOPs) per forward pass and significant memory to store intermediate feature maps. This causes longer inference times and practical need for GPU or hardware acceleration to achieve the real-time performance.

We develop a mathematical framework for evaluating and comparing keypoint-based image matching strategies, formalize the matching pipeline and compare classical and learned methods in terms of inlier correspondences and runtime. To evaluate image matching techniques, we construct a custom dataset consisting of a satellite imagery collected by utilizing the Google Maps Static API (Google, 2025), 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.

1. Scale-Invariant Feature Transform (SIFT)

The SIFT is a hand-crafted pipeline for detecting and describing local features. It constructs the scale space representation $L(\mathbf{x}, \sigma)$ by convolving the image with a variable-scale Gaussian kernel $G(\mathbf{x}, \sigma)$,

$$L(\mathbf{x}, \sigma) = G(\mathbf{x}, \sigma) * I(\mathbf{x}),$$

where $\mathbf{x} = (x, y)^\top$ denotes coordinates of the input image I, $\sigma > 0$ is the scale parameter. To detect scale-invariant keypoints, the SIFT computes the Difference of Gaussians (DoG),

$$D(\mathbf{x}, \sigma) = L(\mathbf{x}, k\sigma) - L(\mathbf{x}, \sigma),$$

where $k > 1$ is the multiplicative factor between successive scales. Local extrema of $D(\mathbf{x}, \sigma)$ in both spatial and scale dimensions are selected as candidate keypoints. Each detected keypoint is assigned with a dominant orientation to achieve rotation invariance: $\theta(\mathbf{x}) = \arctan(L_y/L_x)$, where L_x and L_y denote the image derivatives computed from the smoothed image $L(\mathbf{x}, \sigma)$, to introduce a canonical direction for each keypoint based on the predominant edge direction in its local neighborhood so that the descriptor can be rotated relative to this orientation and remain consistent across rotated views.

The SIFT descriptor is formed by sampling gradient magnitudes and orientations in a region around the keypoint, partitioning this region into a grid of spatial cells, and computing orientation histograms in each cell. The bins number in each histogram determines the final descriptor dimension D, which depends on the cells number and the number of orientation bins per each cell. The resulting vector is normalized to the unit length to improve the robustness.

2. The SuperPoint

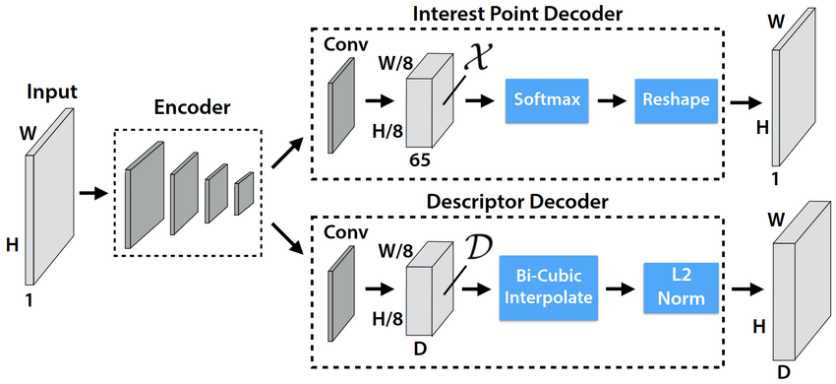

Fig. 1. The SuperPoint (DeTone, Malisiewicz, & Rabinovich, 2018) network architecture. The input image is processed by a shared convolutional encoder into a low-resolution feature map which is then split into two decoder "heads" for the interest point detection and descriptor extraction

The SuperPoint backbone is a fully-convolutional neural network consisting of convolutional layers with the ReLU activations $f(x) = \max(0, x)$ interleaved with three 2×2 max-pooling layers which reduce the spatial resolution by a factor of 8 to $H_c = H/8$ and $W_c = W/8$. The backbone feeds into two decoder heads. The "detector" head, or interest point decoder (Fig. 1, upper branch) outputs a score tensor $\mathbf{z} \in \mathbb{R}^{H_c \times W_c \times 65}$, where the first 64 channels represent positions within an 8×8 grid cell and the 65-th channel ("dustbin") indicates the absence of a keypoint. A channel-wise softmax,

$$\mathbf{p}_i = \text{softmax}(\mathbf{z}_i), \quad p_{i,k} = \frac{e^{z_{i,k}}}{\sum_{u=1}^{65} e^{z_{i,u}}},$$

produces a probability vector $\mathbf{p}_i \in \mathbb{R}^{65}$ for each spatial location i. The detector is first trained on a synthetic dataset where ground-truth keypoints are defined at corners and intersections of primitive geometric figures (e.g., lines, circles, triangles, squares). Neural network training minimizes the spatial cross-entropy in the general definition (Goodfellow et al., 2016),

$$\mathcal{L}_{\text{detector}} = -\frac{1}{H_c W_c} \sum_{i=1}^{H_c W_c} \sum_{k=1}^{65} y_{i,k} \log p_{i,k}, \tag{1}$$

where $y_{i,k} \in \{0, 1\}$ encodes the ground-truth channel for the spatial location i in the $H_c \times W_c$ output grid. Then, in the homographic adaptation phase, the trained detector is applied to multiple random homographic warps of real, unlabeled images, resulting much richer set of interest points than traditional corner detectors – not only corners and intersections but also textured patches, edges and other repeatable patterns. Detections are warped back to form a pseudo ground-truth, and the network is fine-tuned using the same loss (1).

The descriptor head (Fig. 1, lower branch) outputs $\mathbf{f} \in \mathbb{R}^{H_c \times W_c \times D}$, which is bicubically upsampled to $H \times W \times D$ and is L_2-normalized for every pixel. The descriptor training uses the same real images and random homographic warps as in the homographic adaptation stage. Keypoints identified by the detector from the previous training stage are extracted from both the original and warped images, and the known homography is used to find corresponding positions between them. Let $d_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|_2$ be the Euclidean distance between descriptors and $s_{ij} \in \{0, 1\}$ indicate the matches. The descriptor loss is the contrastive loss (Goodfellow et al., 2016):

$$\mathcal{L}_{\text{descriptor}} = \sum_{i=1}^{N} \sum_{j=1}^{M} \lambda s_{ij} \max(0, m_p - d_{ij}) + (1 - s_{ij}) \max(0, d_{ij} - m_n),$$

where m_p and m_n are the positive and negative margins, respectively, and λ balances the contributions of positive and negative terms. The first term penalizes matching pairs ($s_{ij} = 1$) whose distance exceeds the positive margin m_p, encouraging descriptors for corresponding points to be close in feature space. The second term penalizes non-matching pairs ($s_{ij} = 0$) whose distance is less than the negative margin m_n, pushing unrelated descriptors apart. The network parameters for both detector and descriptor heads are optimized jointly using stochastic gradient descent with backpropagation.

3. Comparative Study

To objectively compare feature-based image matching methods, such as SIFT and SuperPoint, we adopt a geometric verification strategy based on the Random Sample Consensus (RANSAC) (Fischler, & Bolles, 1981) inlier statistics. Although learned matchers like SuperGlue (Sarlin et al., 2020) significantly enhance the performance of the SuperPoint, this work isolates the contribution of the feature extractor itself by using a standard RANSAC-based evaluation pipeline, ensuring compatibility with both classical and learned descriptors.

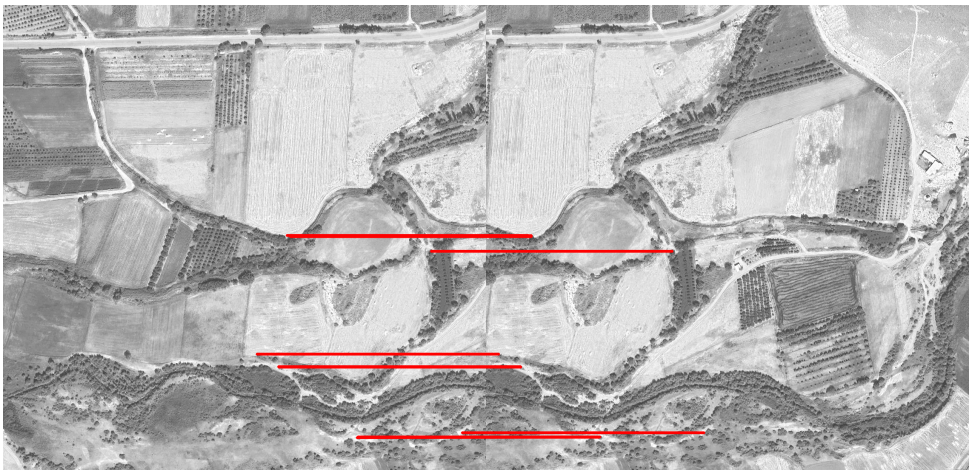

Fig. 2. Keypoint matches between two neighbouring images with an overlapping region. Keypoints are obtained by using the SIFT. The red lines indicate inlier keypoint matches retained by the RANSAC filtering. A collection of the satellite images is retrieved by utilizing the Google Maps Static API (Google, 2025)

Given two images of the same scene under varying viewpoint (Fig. 2), a local feature extractor (e.g., SIFT or SuperPoint) identifies keypoints and their descriptors. The descriptor matching yields the candidate correspondences which are typically noisy and include incorrect matches due to descriptor ambiguity or repeated structures. To filter out incorrect matches, the RANSAC geometric model estimation is applied, estimating a homography matrix **H** by iteratively selecting random minimal samples and maximizing the number of inliers, defined as matches that satisfy the geometric model within a tolerance ε. A match (x_i, x_i') is counted as an inlier if $||x_i' - Hx_i|| < \varepsilon$. This leads to a set of inlier correspondences $I \subset M$, from the full set of candidate matches M. We define the inlier ratio as $r = |I|/|M|$. This ratio quantifies the geometric consistency of matched keypoints. A high inlier ratio indicates that many descriptor matches are geometrically valid and, therefore, reflects higher feature precision and robustness.

Table 1

Comparison of accuracy and computational efficiency. The number next to each method indicates the number of keypoints extracted per image. Measurements are taken on the same hardware setup in CPU-only mode.

Algorithm	Inlier ratio r (overlapping images), %	Inlier ratio r (non-overlapping images), %	Detection & description time, ms per image
SIFT 100	32.21	0.53	111.53
SIFT 200	32.80	0.59	112.83
SIFT 500	34.45	0.42	115.00
SuperPoint 100	38.96	0.02	728.84
SuperPoint 200	41.34	0.04	729.47
SuperPoint 500	43.81	0.08	726.19

Discussion and conclusions

We compared the classical SIFT detector-descriptor pipeline with the learned SuperPoint method in the context of image matching for a vision-guided system. Both methods are evaluated using a custom-constructed dataset of satellite images, with the RANSAC-based geometric verification employed to filter correspondences and compute the inlier ratio as the primary accuracy metric. While the SIFT offers lower computational cost and CPU-friendly performance, the SuperPoint achieves insignificantly higher inlier ratios and improved robustness to viewpoint and illumination changes, at the expense of multiply increased inference time and hardware requirements (Tab. 1). In line with (Bojanić et al., 2019; Prystavka et al., 2025) comparative studies and related evaluations, our findings confirm that classical algorithms retain a practical advantage for real-time navigation tasks and deliver competitive performance, often matching learned approaches. These findings are relevant to navigational systems (Se, Lowe, & Little, 2005; Mur-Artal, Montiel, & Tardos, 2015; Zhu, Liu, & Cai, 2015) and vision-guided vehicles (Hoffmann et al., 2006; Kawamura et al., 2023), where computational resources, energy consumption, and real-time constraints must be balanced.

Future work will extend this analysis to the challenging task of cross-view matching between aerial/satellite imagery, where large scale variations, seasonal changes, and lighting differences present additional difficulties. We also plan to investigate descriptor aggregation methods, e.g. VLAD, to enable efficient large-scale image retrieval, which can benefit both navigation and mapping pipelines.

Sources of funding. This work was supported by the Simons Foundation grant (SFI-PD-Ukraine-00014586, O.S.) and the project 0125U000299 of the National Academy of Sciences of Ukraine.

References

Bojanić, D., Bartol, K., Pribanić, T., Petković, T., Donoso, Y. D., & Mas, J. S. (2019). On the comparison of classic and deep keypoint detector and descriptor methods. In *IEEE 11th symposium on image and signal processing and analysis* (pp. 64–69). https://doi.org/10.1109/ISPA.2019.8868792

Boyun, V., Kasim, A., Voznenko, L., & Matvienko, O. (2025). Three models for creating stereo images from UAV sensor data. *International scientific conference Information technologies and computer modelling*, 111–114 [in Ukrainian]. [Боюн, В., Касім, А., Возненко, Л., & Матвієнко, О. (2025). Три моделі створення стереозображень за сенсорними даними БПЛА. *Міжнародна науково-практична конференція Інформаційні технології та комп'ютерне моделювання*, 111–114].https://journal.comp-sc.if.ua/test/index.php/ITCM/article/view/716

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 224–236). https://doi.org/10.1109/CVPRW.2018.00060

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

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep Learning*. MIT Press.

Google. (2025). *Maps static API (Accessed: 14.08.2025)*. https://developers.google.com/maps/documentation/maps-static/

Hoffmann, F., Nierobisch, T., Seyffarth, T., & Rudolph, G. (2006). Visual servoing with moments of SIFT features. In *IEEE conference on systems, man and cybernetics* (p. 4262-4267). https://doi.org/10.1109/ICSMC.2006.384804

Huang, Q., Guo, X., Wang, Y., Sun, H., & Yang, L. (2024). A survey of feature matching methods. *IET Image Processing*, 18(6), 1385–1410. https://doi.org/10.1049/ipr2.13032

Kawamura, E., Dolph, C., Kannan, K., Brown, N., Lombaerts, T., & Ippolito, C. A. (2023). VSLAM and vision-based approach and landing for advanced air mobility. In *AIAA SciTech 2023 Forum* (p. 2196). https://doi.org/10.2514/6.2023-2196

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5), 1147–1163. https://doi.org/10.1109/TRO.2015.2463671

Prystavka, P., Cholyshkina, O., Kovtun, O., & Pryshchepa, D. (2025). Automation of UAV navigation support based on SIFT-like methods. In *International Workshop on Computational Intelligence (IWSCI)* (pp. 227–239). https://ceur-ws.org/Vol-4035/

Rusyn, B., Lutsyk, O., & Kosarevych, R. (2021). Evaluating the informativity of a training sample for image classification by deep learning methods. *Cybernetics and Systems Analysis*, 57(6), 853–863. https://doi.org/10.1007/s10559-021-00411-4

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In *IEEE conference on computer vision and pattern recognition* (pp. 4938–4947). https://doi.org/10.1109/CVPR42600.2020.00499

Scaramuzza, D., & Fraundorfer, F. (2011). Visual odometry [tutorial]. *IEEE robotics & automation magazine*, 18(4), 80–92. https://doi.org/10.1109/MRA.2011.943233

Se, S., Lowe, D. G., & Little, J. J. (2005). Vision-based global localization and mapping for mobile robots. *IEEE Transactions on robotics*, 21(3), 364–375. https://doi.org/10.1109/TRO.2004.839228

Zhu, Q., Liu, C., & Cai, C. (2015). A novel robot visual homing method based on SIFT features. *Sensors*, 15(10), 26063–26084. https://doi.org/10.3390/s151026063

Отримано редакцією журналу / Received: 26.05.25
Прорецензовано / Revised: 21.09.25
Схвалено до друку / Accepted: 10.10.25

Олег САМОЙЛЕНКО, асп.
ORCID ID: 0000-0003-4466-4004
e-mail: oleh.samoilenko@imath.kiev.ua
Інститут математики НАН України, Київ, Україна

МАТЕМАТИЧНІ ПРОБЛЕМИ ЗІСТАВЛЕННЯ ОЗНАК ДЛЯ ОБ'ЄКТІВ З ОБМЕЖЕНИМИ РЕСУРСАМИ, КЕРОВАНИХ НА ОСНОВІ ЗОРУ

Візуальне сприйняття відіграє ключову роль в автономних транспортних засобах, де зображення слугують основним вхідним сигналом для навігації, побудови карт, оцінки стану та керування рухом. Такі системи часто працюють в умовах реального часу та обмежених ресурсів, що вимагає алгоритмів виявлення та зіставлення ознак, які є точними й обчислювально ефективними. Важливою складовою таких обчислювальних послідовностей є визначення ключових точок – характерних місць зображення $\mathbf{x} \in \mathbb{R}^2$ ***у цифровій функції зображення*** $I : \mathbb{R}^2 \to \mathbb{R}^c$***, де*** c ***– кількість каналів, а*** $I(\mathbf{x}) \in \mathbb{R}^c$ ***– локальна інтенсивність пікселя. Кожна виявлена ключова точка асоціюється з дескриптором*** $\mathbf{d} \in \mathbb{R}^D$***, що кодує локальну візуальну структуру у формі, інваріантній до зсуву, обертання та помірної зміни масштабу. Ключові точки між двома зображеннями*** $\mathbf{x}_i$ ***та*** $\mathbf{x}'_i$ ***зіставлено шляхом порівняння дескрипторів, утворюючи відповідності*** $\mathbf{x}'_i \leftrightarrow \mathbf{x}_i$***. Геометричну верифікацію виконано шляхом оцінки матриці перетворення*** $\mathbf{H} \in \mathbb{R}^{3\times3}$ ***і прийняття*** $|\mathbf{x}'_i - \mathbf{H}\mathbf{x}_i| < \varepsilon$***, де*** ε ***– допустима похибка. Частку таких внутрішніх відповідностей, що називається коефіцієнтом валідних відповідностей, використано як метрику точності. Математично оцінено класичні методи визначення ключових точок, такі як Scale-Invariant Feature Transform (SIFT), у порівнянні з методами на основі машинного навчання, такими як SuperPoint, застосовуючи їх до вручну сформованого набору супутникових зображень. Продуктивність проаналізовано з позиції коефіцієнта валідних відповідностей та обчислювальної ефективності, що відображає компроміс між надійністю й використанням ресурсів. Отримані результати покликані надати рекомендації щодо інтеграції легких, але точних алгоритмів комп'ютерного зору в платформи, такі як БПЛА, забезпечуючи надійну візуальну одометрію та одночасну локалізацію і картографування в умовах обмежених апаратних ресурсів.***

Ключові слова: ***безпілотні літальні апарати, зіставлення зображень, виявлення ознак, комп'ютерний зір, розпізнавання образів.***

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження(у зборі, аналізі чи інтерпретації даних, якщо це мало місце), у написанні рукопису та в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study (in the collection, analyses or interpretation of data if applicable), in the writing of the manuscript as well as in the decision to publish the results.