Electronics and Information Technologies, 2025, 32, 67–86
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.932.2

COMPREHENSIVE SPATIAL-GEOMETRIC EVALUATION OF KEYPOINT DETECTORS

Andriy Fesiuk**, Yuriy Furgala
Faculty of Electronics and Computer Technologies
Ivan Franko National University of Lviv,
50 Drahomanova St., 79005 Lviv, Ukraine

Fesiuk A. V., Furgala Y. M. (2025). Comprehensive Spatial-Geometric Evaluation of Keypoint Detectors. *Electronics and Information Technologies*, 32, 67–86. https://doi.org/10.30970/eli.32.5

ABSTRACT

Background. Local features are essential components of modern computer vision systems, such as SLAM and 3D reconstruction. Traditional evaluation protocols for keypoint detection mainly focus on geometric accuracy and repeatability, often neglecting the spatial structure of the point distribution. This complicates algorithm selection for applications where uniform image coverage and the absence of excessive local clustering are important. This work aims to conduct a comprehensive comparison of keypoint detectors using an extended set of metrics that account for both geometric accuracy and the spatial properties of features.

Materials and Methods. The study was conducted on the HPatches dataset using six detectors: SIFT, SURF, ORB, BRISK, KAZE, and AKAZE. Keypoint filtering and geometric verification of correspondences were performed using USAC. Matching quality was assessed through the geometric metrics MMA, Repeatability, and Verification Ratio. Spatial analysis used the metrics CUI, RI, and SCS. To compare keypoint detection methods, a quality index Q was introduced that integrates geometric and spatial indicators.

Results and Discussion. The study showed that selecting points by response strength significantly improves matching accuracy for SIFT, ORB, and BRISK, but may lead to local redundancy of keypoints. KAZE and AKAZE demonstrated the best overall balance, achieving high accuracy along with more uniform scene coverage. ORB tended to form dense clusters in high-contrast regions, thereby reducing its structural effectiveness, whereas SURF consistently delivered high performance regardless of the keypoint selection strategy.

Conclusion. The proposed evaluation method allows a consistent analysis of the geometric and spatial properties of keypoint detectors. It shows that, for a fixed number of keypoints, the performance of the final method depends not only on the geometric accuracy of matches but also on the features of the spatial point distribution. It was observed that the keypoint selection process, especially response-based selection, systematically affects both geometric and spatial characteristics. The Q quality index combines these aspects into a single metric. It can be used to compare detection methods in scenarios that require both reliable matches and well-balanced scene coverage.

Keywords: feature detection, spatial distribution, geometric metrics, image matching.

INTRODUCTION

Computer vision systems are widely used in navigation, 3D scene reconstruction, visual tracking, and augmented reality [1-5]. In many of these applications, keypoints and their descriptors play a central role in establishing correspondences between images. The

quality of matching directly affects the accuracy of geometric model estimation, tracking stability, and the reliability of subsequent processing stages. Therefore, objective comparison of keypoint detectors remains relevant despite the large body of research in this area.

Traditionally, detector performance is evaluated using repeatability, matching accuracy, the inlier ratio after geometric verification, and the homography estimation error. These measures characterize the stability of feature localization and the geometric correctness of matches. At the same time, the spatial properties of keypoint sets are often considered only to a limited extent: non-uniform image coverage, local clustering, or the presence of “blind zones” can reduce the robustness of geometric estimates even when the classical metrics remain acceptable. Some studies introduce coverage or uniformity indicators; however, they are typically analyzed in isolation and not consistently linked to geometric validation outcomes [6, 7].

An additional practical factor is the keypoint selection strategy when the number of keypoints is limited. Selecting the “strongest” features by detector response can improve matching accuracy, but it may also change the spatial profile of the keypoint set and its structural consistency with the scene. A coherent set of metrics that simultaneously accounts for geometric correctness and spatial-structural properties is therefore essential for a well-grounded comparison of detection methods.

In this study, we propose a comprehensive evaluation scheme for keypoint detectors that combines traditional geometric measures, namely MMA, Repeatability, Verification Ratio, with spatial-structural metrics: Coverage Uniformity Index (CUI), Redundancy Index (RI), and Scene Consistency Score (SCS). For an overall comparison, we introduce a quality index Q that integrates geometric and spatial characteristics. Experiments are performed on the HPatches dataset [8, 9] using detectors such as SIFT [10], SURF [11], KAZE [12], AKAZE [13], ORB [14], and BRISK [15], with geometric correspondence validation utilizing USAC [16].

MATERIALS AND METHODS

The experimental study was conducted on the HPatches dataset [8, 9]. This dataset contains images of planar scenes with varying levels of geometric distortion, for which ground-truth homography matrices are available, enabling precise verification of correspondences. HPatches provides two main types of sequences: viewpoint and illumination. In this work, we analyze the viewpoint sequences. Twelve sequences were selected: apprentices, azzola, busstop, cartooncity, dirtywall, london, posters, samples, sunseason, tabletop, talent, and vitro, which contain enough keypoints for all considered methods. For each image sequence, the original image and five images of the same scene with progressively increasing viewpoint changes were used. The ground-truth transformation between the reference image and each target image is a 3×3 projective homography, which was used for overlap computation and geometric evaluation. For each sequence, "reference image - current viewpoint" pairs were formed. Examples of image sequences are presented in **Fig. 1a**, and the first images of all used sequences are shown in **Fig. 1b**.

The study covers six algorithms for local feature detection and description: the floating-point methods SIFT, SURF, and KAZE, and the binary methods ORB, BRISK, and AKAZE. For each detector, the performance dependence on the number of keypoints is examined, with the number of keypoints ranging from 500 to 4000 in increments of 500. To evaluate how feature selection influences subsequent matching quality, two selection strategies were compared:

- raw-order: Selection of the first N points in the order returned by the detection method.

Fig. 1. Examples from the HPatches dataset: (a) the 'vitro' image sequence; (b) all scenes used in the experiments.

- top-response: Sorting points by detector response strength, then selecting the top N features with the highest response. This simulates a scenario where, under a limited keypoint budget, priority is given to the strongest local structures. The response for each method was calculated as follows [17]:
  - SIFT: The amplitude of the extremum in the Difference of Gaussians (DoG) pyramid, which correlates with contrast.
  - SURF: The determinant of the Hessian matrix used for detecting "blob-like" structures.
  - ORB and BRISK: Metrics based on the Harris corner detector for ORB and on AGAST for BRISK.
  - KAZE and AKAZE: The determinant of the Hessian matrix calculated in a nonlinear scale space, which better preserves object boundaries than traditional Gaussian blurring.

Matching was performed using the Brute-Force method, searching for the two nearest neighbors in descriptor space. Euclidean distance was used for SIFT, SURF, and KAZE descriptors, while Hamming distance was applied for ORB, BRISK, and AKAZE. Preliminary filtering involved Lowe's ratio test with a threshold of 0.75 [10].

Geometric verification was conducted using the USAC_DEFAULT method with a homography model and a 3-pixel threshold [16, 18]. Based on the USAC results, an inlier mask was extracted, allowing the separation of points consistent with the global projection. In this experiment, spatial metrics were computed for all detected points and inliers after filtering, enabling comparison of detector properties before and after geometric verification of correspondences.

The Coverage Uniformity Index quantitatively assessed the uniformity with which points covered the image plane. The image plane was divided into a regular grid of $M = 8 \times 8 = 64$ rectangular cells. For each set of points, the number of keypoints N_i in the i cell was counted, and a normalized distribution was formed.

$$p_i = \frac{N_i}{N}, \tag{1}$$

where N is the total number of points.

The obtained empirical distribution is compared with a perfectly uniform distribution, representing a scenario where each cell contains an equal proportion of keypoints:

$$u_i = \frac{1}{M} \tag{2}$$

The total variation distance measures the deviation between p_i and u_i, and the CUI is defined as:

$$CUI = 1 - \frac{1}{2}\sum_{i=1}^{M} |p_i - u_i|, \quad CUI \in [0, 1]. \tag{3}$$

CUI values close to 1 correspond to almost uniform image coverage, while low values indicate an intense concentration of keypoints in a few cells. Conceptually, this metric is based on the coverage evaluation method proposed in [6], in which the spatial distribution of keypoints complements repeatability in assessing detector quality.

The Redundancy Index was used to measure the local redundancy of keypoints. The need to control redundancy is driven by the fact that traditional detector quality metrics can be biased toward methods that produce spatially overlapping or overly dense detections [19, 20]. Such behavior can result in misleadingly high precision values. Additionally, recent research on redundancy removal techniques has demonstrated that explicitly considering local density and distances between points helps reduce descriptor duplication and improves overall image registration efficiency [20, 21], supporting the goal of the proposed evaluation.

The RI calculation was performed for the set of points $P = (x_i, y_i)_{i-1}^{N}$ on the image of size $W \times H$. To eliminate the dependence on the resolution, the coordinates of the points were normalized to the unit square $[0, 1]^2$:

$$\bar{x} = \frac{x_i}{W}, \quad \bar{y} = \frac{y_i}{H}. \tag{4}$$

Next, for each point $(\bar{x}_i, \bar{y}_i)$, the number of neighbors within a normalized radius *r* was counted. The radius is defined as a fraction of the unit square diagonal:

$$r = p\sqrt{2}, \tag{5}$$

where $p = 0.02$, and the factor $\sqrt{2}$ represents the diagonal of the unit square $[0.1]^2$, giving an interpretable definition of the local neighborhood. The choice of the *p* coefficient is based on its representing the typical size of a descriptor's support region, which helps identify points with strongly correlated vector descriptions due to significant spatial overlap.

Local redundancy was evaluated with the saturation function:

$$c_i = \min\left(\frac{n_i - 1}{K_{ref}}, 1\right), \tag{6}$$

where $K_{ref} = 15$, reflecting the empirical limit of information saturation: a concentration of more than 15 points within a single patch does not enhance the geometric model

estimation, for which 4-8 points are enough, but it only adds unnecessary computational load during the matching process.

The global RI score was calculated by averaging the local values:

$$RI = \frac{1}{N}\sum_{i=1}^{N} c_i\,, \quad RI \in [0,1]. \tag{7}$$

High RI values indicate strong local clustering, while low values suggest sparse and unique point arrangements.

In addition to evaluating spatial uniformity, an essential aspect of the analysis is the detector's ability to adapt to the scene's semantic content, as shown in **Fig. 2**. Traditional methods often exhibit a bias toward certain feature types, such as corners, while neglecting other informative elements in the image; this can lead to data loss when reconstructing scenes with complex geometry. To examine this criterion, an automatic image segmentation procedure was developed to categorize the image into three structural classes: high-contrast corner/texture regions T, contour/edge regions C, and homogeneous regions F. Reference masks are generated from gradient analysis [22-24]. The corner mask T is generated using the Harris detector, followed by Gaussian blur and thresholding at the 97.5th percentile of response. The Canny detector produces the edge mask C and excludes pixels already included in mask T. The mask F consists of pixels with low Sobel gradient magnitude = below the 25th percentile that do not intersect with T and C. This method ensures the mutual exclusivity of these sets, enabling the unique classification of each image region.

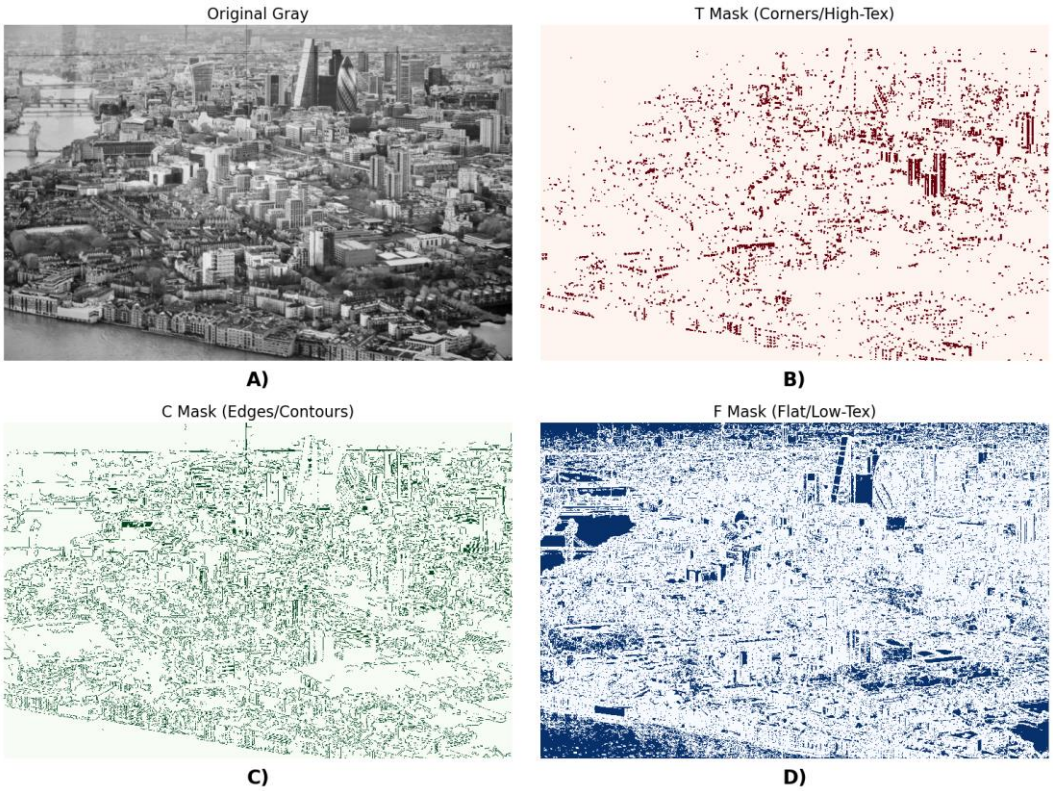

Fig. 2. Structural masks for scene analysis: (A) original image; (B) texture/corner mask T; (C) contour/edge mask C; (D) homogeneous (flat) mask F.

Based on the obtained masks, the relative areas were calculated:

$$\alpha_T = \frac{|T|}{HW}, \quad \alpha_C = \frac{|C|}{HW}, \quad \alpha_F = \frac{|F|}{HW}, \tag{8}$$

where H and W represent the image size, and $|\cdot|$ is the number of pixels in the corresponding mask. Then, the areas were normalized within the structural part of the image:

$$A_{str} = \alpha_T + \alpha_C + \alpha_F, \quad \alpha_T = \frac{\alpha_T}{A_{str}}, \quad \alpha_C = \frac{\alpha_C}{A_{str}}, \quad \alpha_F = \frac{\alpha_F}{A_{str}}. \tag{9}$$

For a given set of keypoints, the number of points falling into each mask (n_T, n_C, n_F) was counted, and then the proportions were calculated:

$$n_{str} = n_T + n_C + n_F, \quad \beta_T = \frac{n_T}{n_{str}}, \quad \beta_C = \frac{n_C}{n_{str}}, \quad \beta_F = \frac{n_F}{n_{str}}. \tag{10}$$

The consistency score of the keypoint distribution with the scene structure was defined as the complement of the total variation distance between α and β [7]:

$$SCS = 1 - \frac{1}{2}(|\beta_T - a_T| + |\beta_C - a_C| + |\beta_F - a_F|), \quad SCS \in [0, 1]. \tag{11}$$

Scene Consistency Score values close to 1 indicate that keypoints appropriately cover structural types in proportion to their actual presence in the scene. In contrast, lower values indicate a systematic bias, such as an excessive concentration on contours or a dominance of detections in texture or corner regions.

To evaluate classical performance metrics, standard protocols used in the HPatches benchmark [8, 9] were applied.

Repeatability was only calculated within the common overlap region of the images. A point x in the first image was considered repeatable if, after being projected through homography $H_{1\to2}$ into the second image, there was a point $\acute{x}$ within $\varepsilon = 3$ pixels of the projected position $n(H_{1\to2}\acute{x})$. To prevent multiple counts of the duplicate detection, "one-to-one" matching was employed [25, 26].

Mean Matching Accuracy was calculated on the set of points that passed Lowe's ratio test [10]. The MMA value was defined as the ratio of correct matches, with a reprojection error $\tau = 3$ in pixels, relative to the Ground Truth homography among all detected pairs [8].

To evaluate the efficiency of keypoint use, the Verification Ratio was employed. Unlike MMA, it is not based on the true homography, but on the results of filtering using the USAC method.

$$VR = \frac{N_{inliers}}{N_{total}}, \tag{12}$$

where $N_{inliers}$ is the number of points consistent with the found geometric model, N_{total} is the total number of detected points.

For a comprehensive comparison of detectors based on geometric and spatial-structural quality, an integral Quality Index (Q) was formulated. Its construction is based on

a weighted linear combination of two aggregated components: the geometric component *G* and the structural-spatial component *S*.

The geometric component *G* characterizes the accuracy of correspondence establishment. It is defined as the arithmetic mean of the normalized values of repeatability, matching accuracy, and verification ratio obtained after the filtering stage:

$$G = \frac{1}{3}(MMA_{\tau=3} + Rep_{\varepsilon=3} + VR). \tag{13}$$

The structural-spatial component S evaluates scene coverage quality and the detector's adaptation to scene content by combining metrics such as structural consistency, distribution uniformity, and local redundancy:

$$S = \frac{1}{3}\left(CUI_{filt} + (1 - RI_{filt})^2 + SCS_{filt}\right), \tag{14}$$

where $(1 - RI_{filt})^2$ indicates a preference for low local redundancy on a $[0, 1]^2$ scale, and its quadratic form emphasizes differences between sets with moderate and high local redundancy while also diminishing the influence of small RI values in the low redundancy range.

The weighting coefficients were assigned according to the golden ratio principle [27]. Since geometric accuracy is essential for most computer vision tasks, it was assigned a higher weight of 0.62, whereas structural-spatial quality is an additional important factor with a weight of 0.38.

$$Q = 0.62 \cdot G + 0.38 \cdot S. \tag{15}$$

The validation of the quality index is performed by analyzing the correlation between the *Q* values and the first principal component *PC1*, obtained via *PCA* [28] for all metrics, which allowed verifying the consistency of the proposed integral indicator with the multidimensional data structure.

The standard HPatches protocol is used for Repeatability, MMA, and Verification Ratio [8,9]. CUI adopts the coverage concept from [6] in a grid-based form. RI, SCS, and the Quality Index are introduced in this study. Masks are computed using classical Harris, Sobel, and Canny operators [22-24], whereas T/C/F partitioning and normalization are defined in this work.

RESULTS AND DISCUSSION

The analysis of the Coverage Uniformity Index in **Fig. 3** shows that, for all detectors, coverage uniformity improves as the quantity of keypoint *N* increases. After USAC geometric validation, CUI_{filt} values are consistently lower than CUI_{raw} because inliers form a more selective subset of correspondences. For KAZE and AKAZE, top_response provides noticeably more uniform coverage even at low keypoint counts; as *N* increases, both modes converge, reaching high CUI_{raw} values of about 0.65-0.70 at *N* = 4000. After filtering, these two methods maintain some of the best CUI_{filt} values among the detectors, around 0.40-0.45 for large *N*.

For SIFT and BRISK, selecting the strategy has the most significant impact: switching to top_response significantly increases CUI, even for small *N*. For SIFT at *N* = 500, CUI_{raw} the value rises from 0.16 to 0.47, and this improvement persists across the entire keypoint range. After USAC filtering, CUI values decrease, but the difference between methods

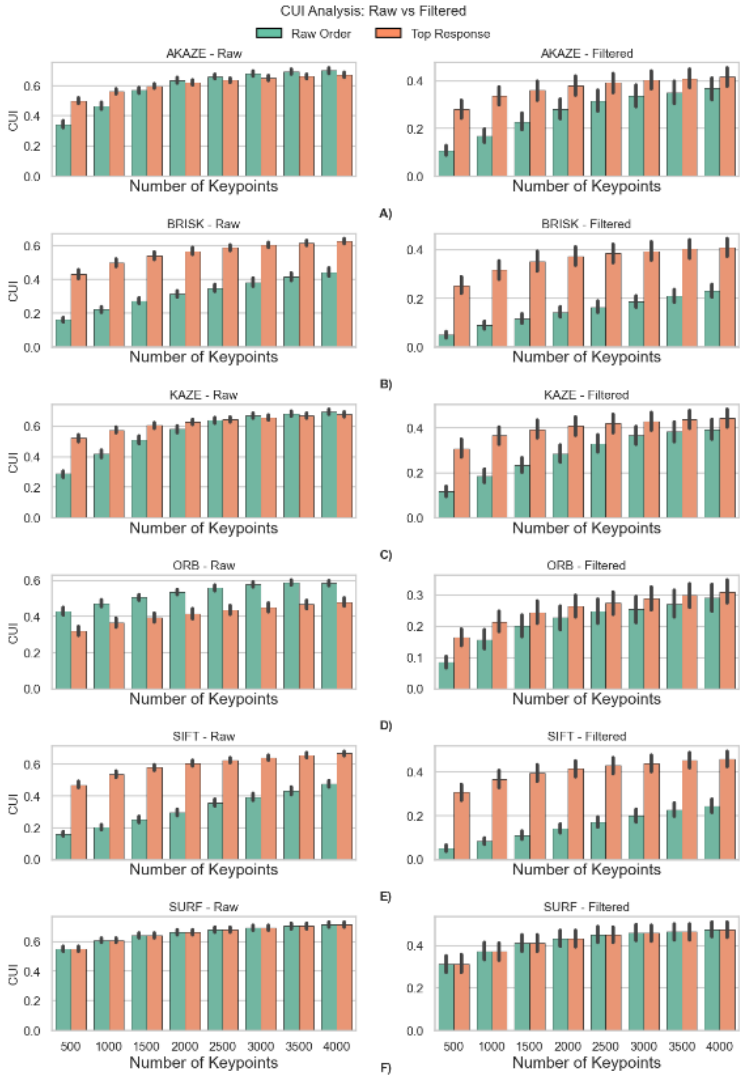

Fig. 3. Average CUI vs. the number of keypoints for raw sets (left) and filtered (right) across detectors under two selection strategies: A) AKAZE; B) BRISK; C) KAZE; D) ORB; E) SIFT; F) SURF.

remains steady, showing a more uniform spatial distribution of inliers with the top_response approach.

A different pattern appears for ORB: on unfiltered points, the raw_order mode yields a higher CUI, but after filtering, the difference between modes diminishes and shifts in favor of top_response. SURF shows the most consistent behavior: CUI is independent of the selection mode, indicating that the detector returns values already sorted by top_response, with CUI_{raw} values staying high at around 0.6 or above for all *N*. After filtering, SURF also maintains some of the highest CUI_{filt} values, approximately 0.47 at *N* = 4000. In conclusion, SURF offers the best average uniform coverage, followed by KAZE and AKAZE. SIFT and BRISK rank next, for which the top_response mode is essential for achieving high CUI.

The results of the Redundancy Index metric in **Fig. 4** show that as the number of keypoints increases, local redundancy tends to grow for most detectors in the unfiltered set RI_{raw}. This indicates a tendency for points to densify within the "strongest" local structures.

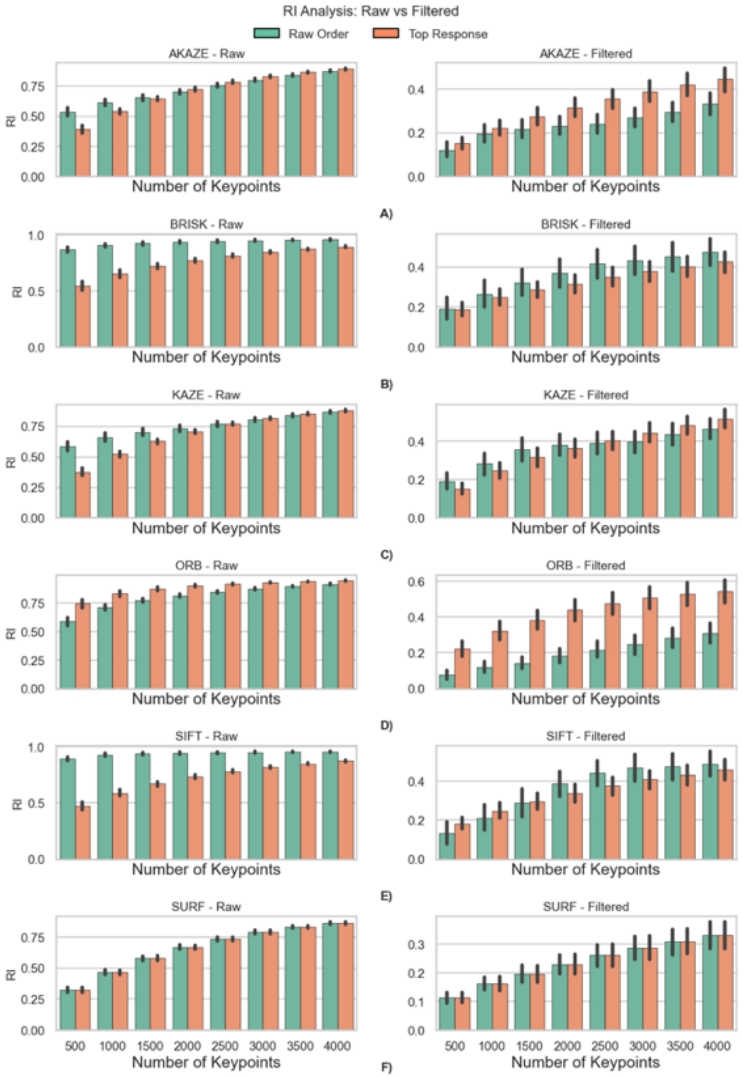

Fig. 4. Average RI vs. the number of keypoints for raw sets (left) and filtered (right) across detectors under two selection strategies: A) AKAZE; B) BRISK; C) KAZE; D) ORB; E) SIFT; F) SURF.

The highest RI_{raw} values are observed for SIFT and BRISK in raw_order mode, averaging approximately 0.94 and 0.93 across the entire range, which suggests pronounced clustering. Switching to top_response for these methods significantly changes RI values, reducing redundancy to around 0.72 for SIFT and 0.77 for BRISK. This indicates that selecting keypoints based on response produces a “sparser” point distribution.

For AKAZE, KAZE, and SURF, the unfiltered RI_{raw} values are lower, generally ranging from 0.65 to 0.75. The effect of top_response is weaker and depends on the method: for KAZE, changes are minimal, while for AKAZE, selecting by response slightly decreases redundancy. In contrast, for ORB, the opposite trend occurs: top_response increases local clustering even at the RI_{raw} level, averaging about 0.89 versus 0.80 for raw_order, which is consistent with ORB's tendency to concentrate points in a limited set of corner or high-contrast regions.

After filtering, RI_{filt} decreases significantly and shifts to a range characteristic of moderate inlier density, mainly between 0.2 and 0.4, indicating the selection of more

structurally consistent and spatially "cleaned" matches. AKAZE and SURF show the lowest inlier values, approximately 0.24 for raw_order and 0.24 for SURF in both modes. For AKAZE, top_response increases RI_{filt} by about 0.32, indicating that inliers remain relatively sparse. Conversely, for ORB, the top_response mode substantially increases inlier redundancy: RI_{filt} rises from approximately 0.20 to 0.43, indicating denser local clusters even after geometric verification. Overall, the RI results suggest that the keypoint selection strategy can significantly alter the local structure of the set: for SIFT and BRISK, top_response reduces redundancy, whereas for ORB it enhances clustering, particularly at the inlier level.

Fig. 5 shows the dependencies of the Scene Consistency Score. For the selected HPatches scenes, the average area proportions of structural zones $a_T = 0.083$, $a_T = 0.278$, and $a_F = 0.639$, indicating that low-texture regions are predominant on average. Under these conditions, SURF exhibits the highest consistency in the inlier distribution

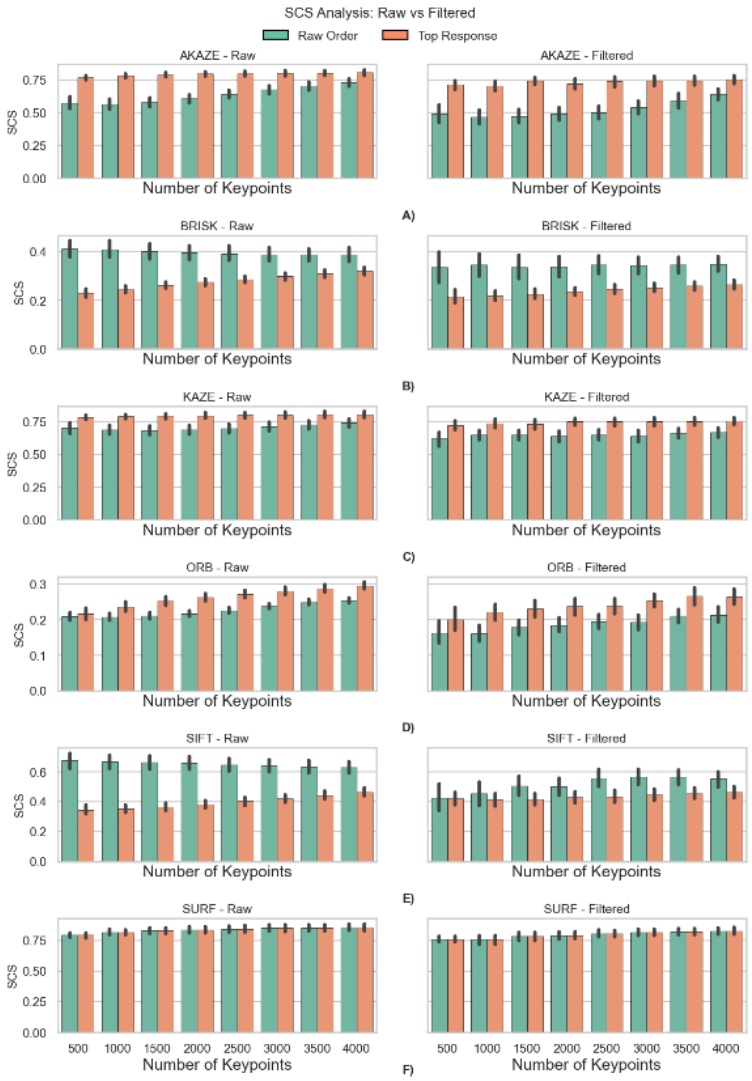

Fig. 5. Average SCS vs. the number of keypoints for raw sets (left) and filtered (right) across detectors under two selection strategies: A) AKAZE; B) BRISK; C) KAZE; D) ORB; E) SIFT; F) SURF.

aligned with the scene structure: SCS values remain high, at approximately 0.83 before filtering and 0.79 after filtering. High values are also typical of KAZE and AKAZE, especially in the top_response mode, where SCS_{filt} stays around 0.73-0.74.

In contrast, ORB and BRISK have the lowest SCS values, averaging 0.3 and 0.4 before filtering, and 0.2-0.25 after USAC processing, indicating a significant discrepancy between scene structure and the actual inlier distribution. For SIFT, a pronounced dependence on the selection strategy is observed: in raw_order, SCS_{raw} it remains relatively high at approximately 0.65, whereas top_response reduces consistency to approximately 0.4, and the advantage of raw_order persists after filtering.

The interpretation of these differences is supported by the structural analysis in **Fig. 6**, where the proportions β_T, β_C, β_F are compared before and after filtering and for the two keypoint selection strategies. For ORB and BRISK, inliers are sharply skewed toward corner/texture zones: β_T reaches approximately 0.8-0.9, while the contribution of

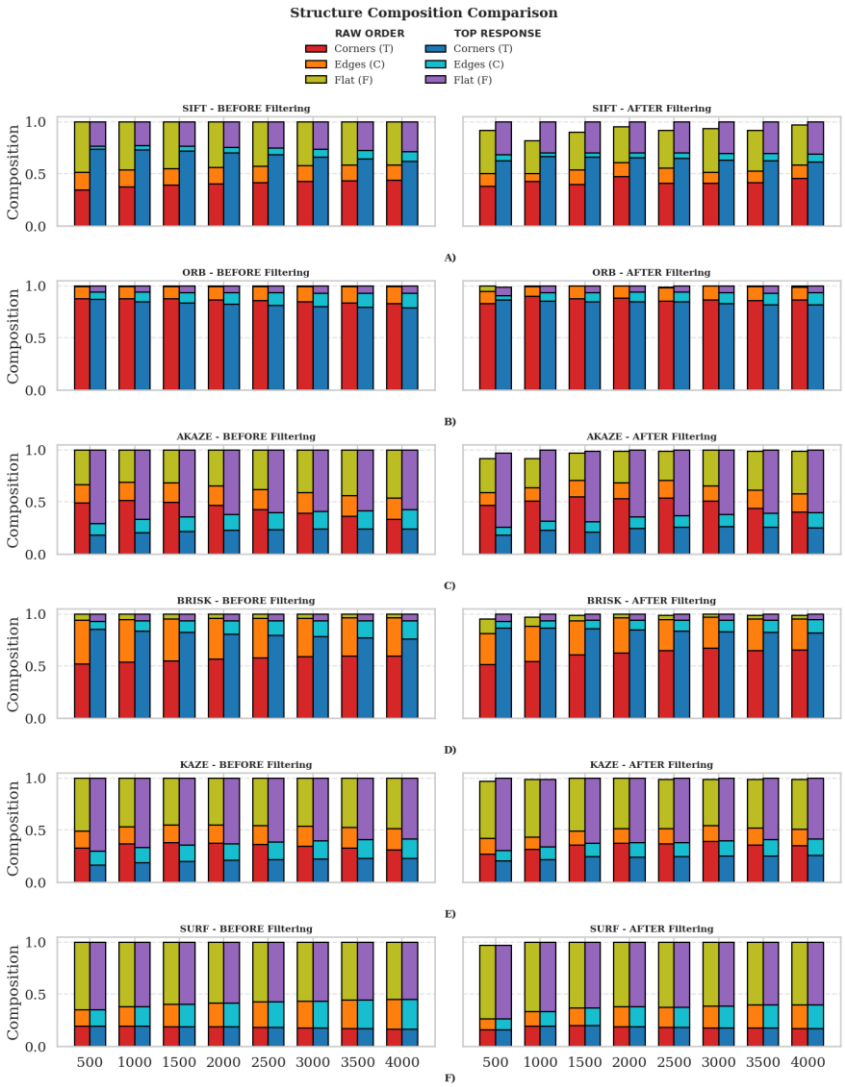

Fig. 6. Average Structure Composition vs. the number of keypoints for raw sets (left) and filtered (right) across detectors under two selection strategies: A) SIFT; B) ORB; C) AKAZE; D) BRISK; E) KAZE; F) SURF.

homogeneous regions β_F is minimal; this effect is more potent in top_response, and for ORB, it persists even after USAC processing. For SIFT, the top_response mode also causes a shift toward corner structures, which is consistent with the decline in SCS values. In contrast, KAZE, AKAZE, and SURF produce values with a significant proportion of low-texture regions β_F and moderate contributions from β_T and β_C, with their proportions changing only slightly after filtering. It is precisely this stability and structural "neutrality" that corresponds to the high SCS values observed across a wide range of keypoint counts.

Fig. 7 presents the MMA values, which indicate the percentage of matches that agree with the Ground Truth homography within a 3-pixel threshold. Two main patterns are observed across most detectors: in raw_order mode, MMA values increase with the number of keypoints, and switching to top_response consistently yields higher MMA values at lower keypoint counts than raw_order.

SIFT and ORB show the most substantial reliance on top_response. Even at *N* = 500, MMA scores increase from 0.27 to 0.75 for SIFT and from 0.41 to 0.66 for ORB, demonstrating that selecting by response significantly boosts the number of geometrically correct matches at low keypoint counts. When averaged across all *N*, this is reflected in an increase in MMA from approximately 0.55 to 0.75 for SIFT and from 0.54 to 0.71 for ORB. For AKAZE, KAZE, and BRISK, this effect is also present but less marked. Their mean values climb from roughly 0.64 to 0.76 for AKAZE, from 0.71 to 0.76 for KAZE, and from 0.72 to 0.80 for BRISK. The MMA scores for SURF stay around the 0.7 range.

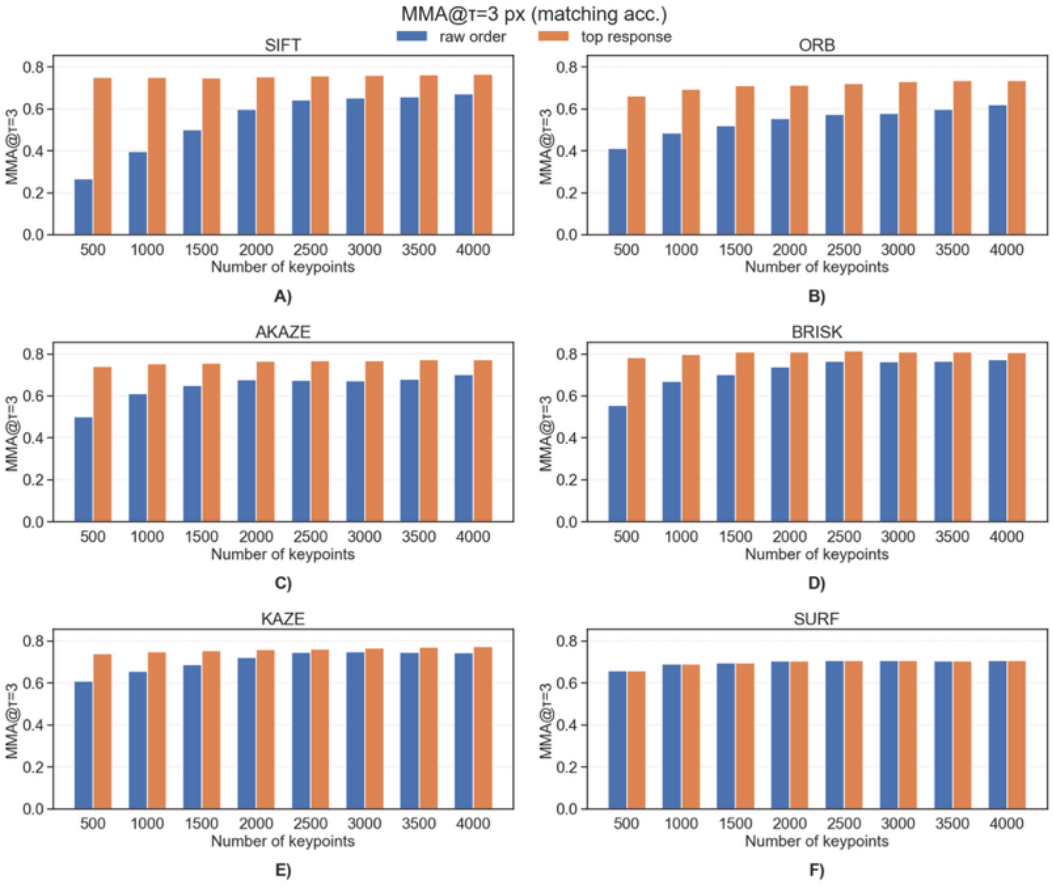

Fig. 7. Average MMA vs. the number of keypoints across detectors under two selection strategies: A) SIFT; B) ORB; C) AKAZE; D) BRISK; E) KAZE; F) SURF.

Overall, the MMA results confirm that selecting keypoints by response enhances the geometric correctness of matches across all methods, with SIFT and ORB showing the greatest sensitivity to the selection strategy.

Fig. 8 shows the Repeatability values, which measure the proportion of keypoints with a matching detection in the paired image within a 3-pixel threshold in the overlap region. Across all detectors, repeatability increases with the number of keypoints; most methods show the most significant improvements at low *N* values, then gradually level off.

KAZE exhibits the most consistent behavior, maintaining high repeatability levels around 0.55-0.60 across nearly all values of *N*, with minimal dependence on the point selection method. This suggests strong keypoint repeatability in viewpoint change tasks. BRISK and AKAZE achieve similar performance levels, although the effect of the selection strategy is more pronounced for BRISK. Switching to top_response consistently boosts repeatability, particularly at lower keypoint counts, reaching about 0.58-0.62 at larger *N*. SURF shows moderate yet highly stable repeatability, roughly in the range of 0.48-0.52.

In raw order mode, SIFT exhibits lower Repeatability, with values ranging from 0.25 to 0.38 across the entire set of keypoints. Nevertheless, selection based on the top response consistently improves the metric, raising it to approximately 0.42-0.50. Regarding ORB, the influence of top response is most significant at low keypoint counts, with a noticeable increase between *N*=500 and *N*=1500. Conversely, at higher keypoint counts, the disparity between the modes diminishes substantially, with both curves converging to approximately 0.57-0.59.

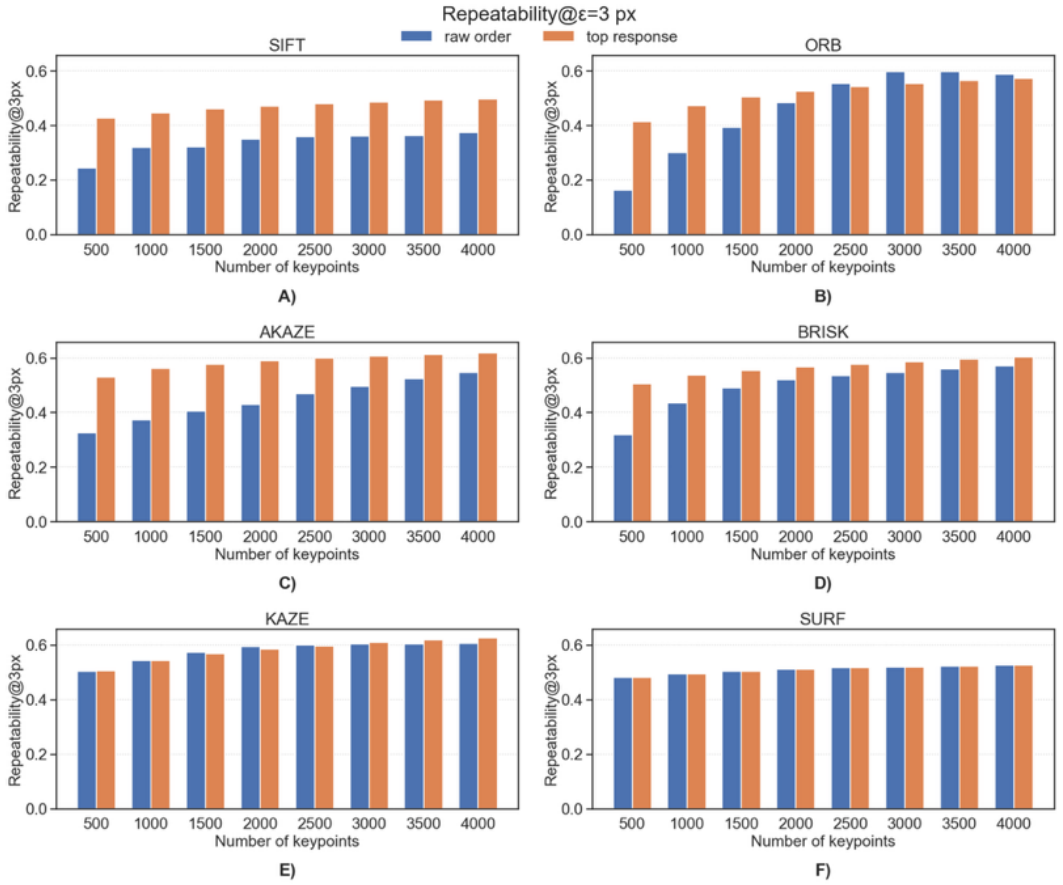

Fig. 8. Average Repeatability vs. the number of keypoints across detectors under two selection strategies: A) SIFT; B) ORB; C) AKAZE; D) BRISK; E) KAZE; F) SURF.

Overall, KAZE demonstrates the highest repeatability. Additionally, top_response primarily improves repeatability for SIFT, AKAZE, and BRISK at low keypoint counts, without altering the overall ranking of the methods.

Fig. 9 shows the dependencies of the Verification Ratio (VR). For most detectors, VR rises with the number of keypoints in the raw_order mode, while top_response consistently yields higher values even at small *N*, causing the curves to saturate early.

The highest VR values are observed for KAZE and SIFT in the top_response mode, at approximately 0.21-0.23. AKAZE also exhibits a comparable top_response of approximately 0.2, whereas in raw_order, the values are noticeably lower. For BRISK, the effect of selection by response is particularly pronounced at small *N* values: VR increases from about 0.09 to 0.17, after which the difference between the selection methods gradually decreases as *N* increases. ORB remains the least effective according to the VR metric; even in top_response, values are around 0.11–0.12, indicating a smaller proportion of points supporting a common homography under the conditions of various HPatches scenes. SURF stands out, with VR values remaining relatively high, though they show a moderate downward trend as the number of keypoints increases, from approximately 0.20 to 0.17. Overall, the best results under top_response are achieved by KAZE and AKAZE, whereas ORB exhibits the lowest proportion of geometrically verified points.

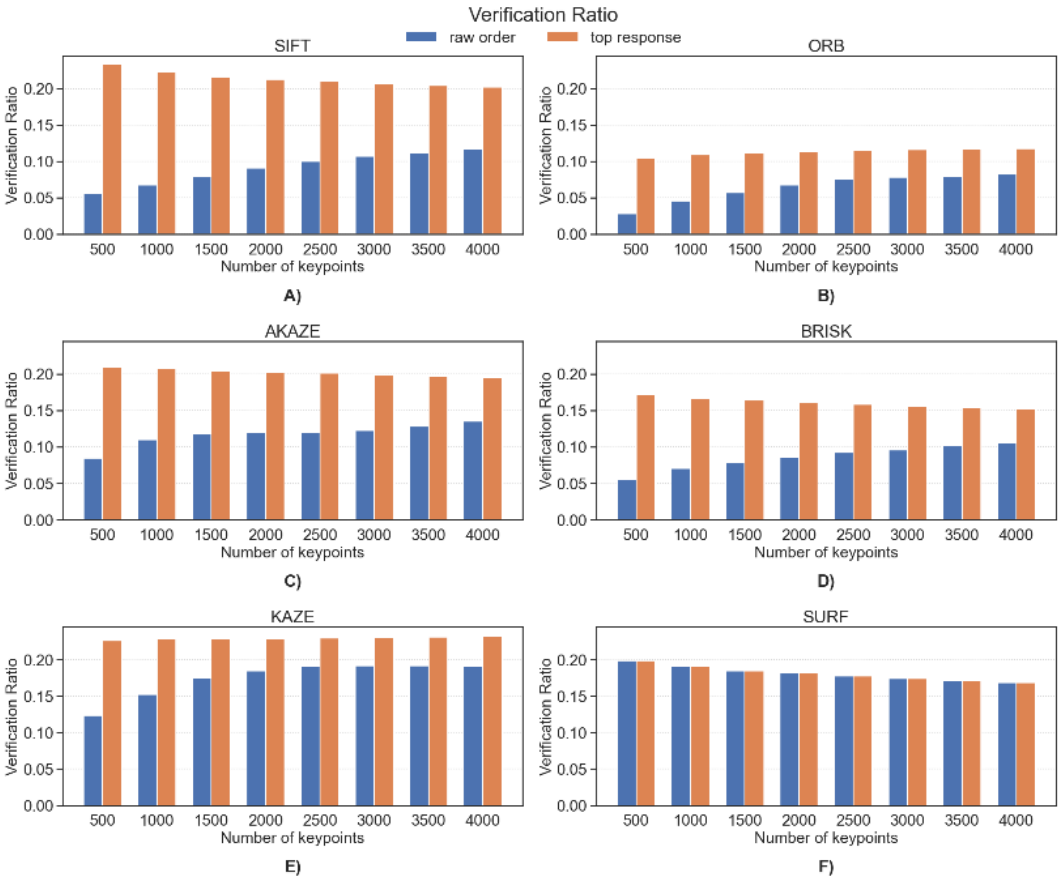

Fig. 9. Average Verification Ratio vs. the number of keypoints across detectors under two selection strategies: A) SIFT; B) ORB; C) AKAZE; D) BRISK; E) KAZE; F) SURF.

Fig. 10 illustrates the dependencies of the Quality Index, which includes the geometric and spatial components. The overall trend is moderate: the influence of the number of keypoints on the Q value is less significant than that of the detector choice and the point selection strategy.

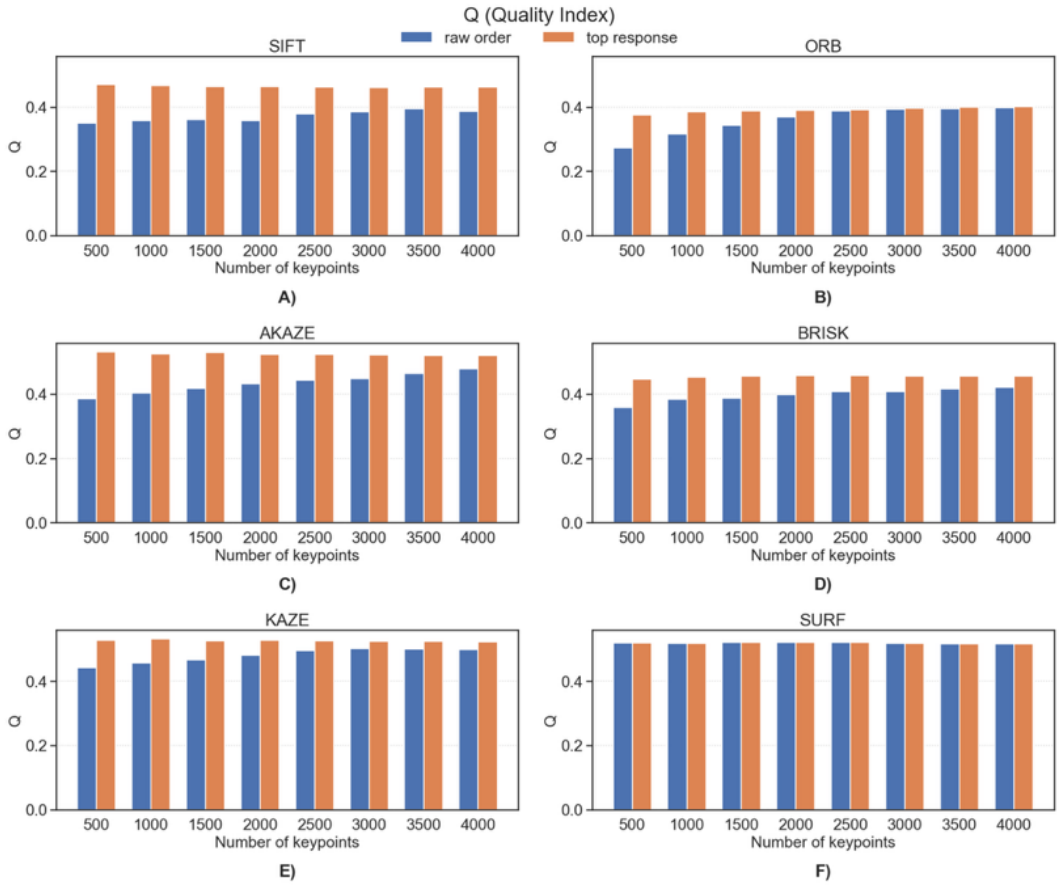

Fig. 10. Average Quality Index vs. the number of keypoints across detectors under two selection strategies: A) SIFT; B) ORB; C) AKAZE; D) BRISK; E) KAZE; F) SURF.

The highest Q values are achieved by the KAZE and AKAZE methods in the top_response mode, with averages of 0.526 and 0.524, respectively. The next best method is SURF, with an accuracy of approximately 0.51. For SIFT, the transition from raw_order to top_response yields a substantial increase in Q, from approximately 0.37 to 0.46, consistent with the simultaneous improvement in geometric and spatial characteristics in this mode. BRISK achieves an intermediate quality level of approximately 0.4 for raw_order and 0.45 for top_response. In contrast, ORB has the lowest Q values of approximately 0.36 and 0.39, respectively, reflecting limitations in both the proportion of geometrically verified correspondences and instructural-spatial metrics.

To verify the consistency of the Q, Principal Component Analysis (PCA) [28] was conducted on the MMA, Repeatability, VR, CUI_{filt}, SCS_{filt}, and RI_{filt} metrics. The first principal component, PC1, accounted for approximately 56% of the total variance and had positive loadings of similar magnitude across all measures, with the largest contributions from VR, CUI, and MMA. The correlation between Q and PC1 was approximatelyFinal_5_Article_text $r = 0.91$, suggesting that the proposed index effectively captures the main latent "axis" of quality derived by the principal component

method. This further confirms the correctness of the selected metrics and their weights in the *Q* formula.

The results obtained show that evaluating detectors solely on standard geometric metrics does not always capture differences in the spatial organization of keypoints, which significantly affect the suitability of the correspondence set. Spatial metrics reveal additional patterns: an increase in geometric performance may be accompanied by changes in coverage uniformity, local redundancy, and the structural consistency of keypoints with the scene; therefore, for a comprehensive comparison, it is advisable to consider both groups of characteristics together. This focus on spatial structure aligns with modern approaches that examine keypoint stability and quality through clustering characteristics and spatial patterns; relevant studies highlight the importance of assessing detector properties beyond matching accuracy [19-21, 29].

The analysis of point selection strategies is also fundamental [30]. An exception is the SURF detector, where both strategies yielded nearly identical results. This occurs due to the specifics of the OpenCV implementation, which returns keypoints already sorted by the response magnitude linked to the Hessian measure 17].

The proposed metric *Q* enables comparison of detectors while accounting for the factors mentioned earlier. PCA was used as an initial validation tool and verified that the *Q* index aligns with the dominant direction of variability in the normalized metrics. A more detailed analysis of the component structure and loading stability is planned for future work.

CONCLUSION

This study offers a comprehensive evaluation of the SIFT, SURF, ORB, BRISK, KAZE, and AKAZE keypoint detectors on the HPatches dataset, using USAC to validate matches geometrically. It shows that assessing methods solely based on traditional geometric metrics - such as MMA, Repeatability, and Verification Ratio - does not fully capture the differences among detectors because it overlooks the spatial arrangement of keypoints. The proposed spatial-structural metrics, CUI, RI, and SCS, provide a quantitative description of frame coverage uniformity, local redundancy, and the consistency of keypoint distribution with scene structure, thereby enhancing standard accuracy analyses.

The results showed that the selection strategy had a greater impact on detector behavior than the number of keypoints. Selecting keypoints by response strength consistently increased MMA, Repeatability, and Verification Ratio across all methods, with the strongest effects observed for SIFT, ORB, and BRISK. At the same time, the analysis of spatial characteristics demonstrated that not all detectors consistently achieve uniform scene coverage and structural consistency. The most balanced CUI, RI, and SCS values were recorded for KAZE and AKAZE; high values were also obtained for SURF, whereas ORB and BRISK were found to be more specialized toward corner-like structures and tended to form redundant, clustered point sets.

The proposed Quality Index, which combines spatial and geometric metrics, enables the generalization of evaluation and comparison across detection methods. The KAZE and AKAZE detectors achieved the highest average *Q* values, with SURF ranked second. The high scores of these methods simultaneously ensure acceptable geometric accuracy, sufficient repeatability, and a spatial distribution of keypoints that is close to uniform. The performance of SIFT and BRISK was competitive in terms of *Q* values when keypoints were selected by response strength. The ORB method remained the least balanced according to the generalized *Q* indicator, despite significant improvements in matching accuracy after sorting keypoints by response.

Principal component analysis indicated that the first component effectively summarized the variability of the metrics and was strongly correlated with the Q index,

thereby providing additional validation of the chosen approach. This result suggests that the proposed indicator accurately aggregates inlier-set properties and can serve as a general criterion for detector comparison in computer-vision tasks where both homography accuracy and scene coverage are essential. In future work, the conclusions should be verified on additional datasets and geometric models – notably the fundamental matrix – and the analysis should be extended to a broader range of imaging conditions, including illumination changes.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The author(s) received no financial support for the research, writing, and/or publication of this article.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, [A.F., Yu.F.]; methodology, [A.F., Yu.F.]; validation, [A.F., Yu.F.]; writing – original draft preparation, [A.F.]; writing – review and editing, [A.F., Yu.F.]; supervision, [Yu.F.].

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Zhou, L., Wu, G., Zuo, Y., Chen, X., & Hu, H. (2024). A comprehensive review of vision-based 3D reconstruction methods. *Sensors, 24*(7), 2314. https://doi.org/10.3390/s24072314

[2] Ye, Z., Bao, C., Zhou, X., Liu, H., Bao, H., & Zhang, G. (2023). EC-SfM: Efficient covisibility-based structure-from-motion for both sequential and unordered images. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2023.3285479

[3] 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

[4] Herrera-Granda, E. P., Berrones-González, A., & Aguilar, W. (2024). Monocular visual SLAM, visual odometry, and structure from motion: A review. *Heliyon, 10*(9), e37356. https://doi.org/10.1016/j.heliyon.2024.e37356

[5] Al-Tawil, Y., Zagraoui, A., & Zeghbib, A.-R. (2024). A review of visual SLAM for robotics: evolution, properties, and future applications. *Frontiers in Robotics and AI, 11*, 1347985. https://doi.org/10.3389/frobt.2024.1347985

[6] Ehsan, S., Kanwal, N., Clark, A. F., & McDonald-Maier, K. D. (2011). Measuring the coverage of interest point detectors. In M. Kamel & A. Campilho (Eds.), *Image Analysis and Recognition (ICIAR 2011)* (Lecture Notes in Computer Science, Vol. 6753, pp. 253-261). Springer. https://doi.org/10.1007/978-3-642-21593-3_26

[7] Mousavi, V., Varshosaz, M., & Remondino, F. (2021). Using information content to select keypoints for UAV image matching. *Remote Sensing, 13*(7), 1302. https://www.mdpi.com/2072-4292/13/7/1302

[8] Balntas, V., Lenc, K., Vedaldi, A., & Mikolajczyk, K. (2017). HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2017.410

[9] Balntas, V., Lenc, K., Vedaldi, A., Tuytelaars, T., Matas, J., & Mikolajczyk, K. (2020). H-Patches: A benchmark and evaluation of handcrafted and learned local

descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(11), 2825-2841. https://doi.org/10.1109/TPAMI.2019.2915233
[10] 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
[11] Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding, 110*(3), 346-359. https://doi.org/10.1016/j.cviu.2007.09.014
[12] Alcantarilla, P. F., Bartoli, A., & Davison, A. J. (2012). KAZE features. In *ECCV 2012* (LNCS 7577, pp. 214-227). https://doi.org/10.1007/978-3-642-33783-3_16
[13] Alcantarilla, P. F., Nuevo, J., & Bartoli, A. (2013). Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *BMVC 2013* (pp. 1-11). https://doi.org/10.5244/C.27.13
[14] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *ICCV 2011* (pp. 2564-2571). https://doi.org/10.1109/ICCV.2011.6126544
[15] Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. In *ICCV 2011* (pp. 2548-2555). https://doi.org/10.1109/ICCV.2011.6126542
[16] M. Ivashechkin, D. Baráth, J. Matas, "USACv20: Robust Essential, Fundamental and Homography Matrix Estimation," 2021. https://doi.org/10.48550/arXiv.2104.05044
[17] Howse, Joseph, and Joe Minichino. “Learning OpenCV 4 Computer Vision with Python 3: Get to grips with tools, techniques, and algorithms for computer vision and machine learning.”, *Packt Publishing Ltd*, 2020.
[18] Fesiuk, A., & Furgala, Y. (2025). Keypoint matches filtering in computer vision: Comparative analysis of RANSAC and USAC variants. *International Journal of Computing, 24*(2), 343-350. https://doi.org/10.47839/ijc.24.2.4018
[19] Hossein-Nejad, Z., & Nasri, M. (2017). RKEM: Redundant keypoint elimination method in image registration. *IET Image Processing, 11*, 273-284. https://doi.org/10.1049/iet-ipr.2016.0440
[20] Hossein-Nejad, Z., Agahi, H., & Mahmoodzadeh, A. (2021). Image matching based on the adaptive redundant keypoint elimination method in the SIFT algorithm. *Pattern Analysis and Applications, 24*(2), 669-683. https://doi.org/10.1007/s10044-020-00938-w
[21] Hossein-Nejad, Z., & Nasri, M. (2022). Clustered redundant keypoint elimination method for image mosaicing using a new Gaussian-weighted blending algorithm. *The Visual Computer, 38*(6), 1991-2007. https://doi.org/10.1007/s00371-021-02261-9
[22] Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference* (pp. 147-151). https://doi.org/10.5244/C.2.23
[23] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8*(6), 679-698. https://doi.org/10.1109/TPAMI.1986.4767851
[24] Soille, P. (2004). *Morphological Image Analysis: Principles and Applications* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-05088-0
[25] Rey-Otero, I., Delbracio, M., & Morel, J.-M. (2015). Comparing feature detectors: A bias in the repeatability criteria. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 3024-3028). https://doi.org/10.1109/ICIP.2015.7351358
[26] Rey-Otero, I., & Delbracio, M. (2015). Is repeatability an unbiased criterion for ranking feature detectors? *SIAM Journal on Imaging Sciences, 8*(4), 2558-2580. https://doi.org/10.1137/15M1007732

[27] Avriel, M., & Wilde, D. J. (1968). Golden block search for the maximum of unimodal functions. *Management Science, 14*(5), 307-319 https://doi.org/10.1287/mnsc.14.5.307
[28] Boudt, K., d'Errico, M., Luu, H. A., & Pietrelli, R. (2022). Interpretability of Composite Indicators Based on Principal Components. *Journal of Probability and Statistics*, 2022**,** 4155384, 1-12. https://doi.org/10.1155/2022/4155384
[29] Mukherjee, S., Lagache, T., & Olivo-Marin, J.-C. (2021). Evaluating the stability of spatial keypoints via cluster core correspondence index. *IEEE Transactions on Image Processing, 30*, 386-401. https://doi.org/10.1109/TIP.2020.3036759
[30] A. Fesiuk and Y. Furgala, (2025). Comparative Study of Feature Detectors and Keypoint Filters in Image Matching. *Electronics and Information Technologies*, (31), (pp. 71-88). http://dx.doi.org/10.30970/eli.31.7

КОМПЛЕКСНЕ ПРОСТОРОВО-ГЕОМЕТРИЧНЕ ОЦІНЮВАННЯ ДЕТЕКТОРІВ КЛЮЧОВИХ ТОЧОК

Андрій Фесюк, **Юрій Фургала***
Факультет електроніки та комп'ютерних технологій
Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 Львів, Україна

АНОТАЦІЯ

Вступ. Локальні ознаки є важливими компонентами сучасних систем комп'ютерного зору, таких як SLAM та 3D-реконструкція. Традиційні підходи до оцінювання методів детекції зосереджуються переважно на геометричній точності та повторюваності, часто не враховуючи просторову структуру розподілу точок. Це ускладнює вибір алгоритму для задач, де важливими є рівномірність покриття кадру та відсутність надлишкової локальної кластеризації. Метою роботи є комплексне порівняння детекторів із використанням розширеного набору метрик, що враховують як геометричну коректність, так і просторові властивості ознак.

Матеріали та методи. Дослідження проведено на наборі даних HPatches для шести детекторів: SIFT, SURF, ORB, BRISK, KAZE, AKAZE. Фільтрація особливих точок виконана методом USAC. Якість зіставлення оцінено за геометричними метриками MMA, Repeatability, Verification Ratio. Для просторового аналізу використано метрики CUI, RI та SCS. Для узагальненого порівняння методів детекції ключових точок запропоновано індекс якості Q, який об'єднує геометричні та просторові показники.

Результати. Результати показали, що стратегія відбору точок за силою відгуку суттєво підвищує точність зіставлення для SIFT, ORB та BRISK, проте призводить до локальної надлишковості ключових точок. Результати методів KAZE та AKAZE продемонстрували найкращий баланс, забезпечуючи високу точність при рівномірному покритті сцени. Метод ORB виявився схильним до формування щільних скупчень у контрастних зонах, що знижує його структурну ефективність, тоді як SURF показав стабільно високі результати незалежно від стратегії відбору ключових точок.

Висновки. Запропонований підхід до оцінювання забезпечує узгоджений аналіз геометричних і просторових властивостей детекторів ключових точок та показує, що за фіксованої кількості ключових точок на підсумкову якість методу істотно впливають не лише показники геометричної коректності відповідностей, а й характеристики просторового розподілу точок. Встановлено, що спосіб вибору особливих точок,

зокрема відбір точок за силою відгуку, систематично змінює як геометричні, так і просторові властивості. Індекс якості Q узагальнює ці аспекти в єдиному показнику та може застосовуватися для порівняння методів детекції у сценаріях, де потрібні одночасно надійні відповідності та збалансоване покриття сцени.

Ключові слова: виявлення ознак, просторовий розподіл, геометричні метрики, співпадіння.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
10 December, 2025	17 December, 2025	17 December, 2025	25 December, 2025