Electronics and Information Technologies, 2025, 31, 71–88
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.89 + 004.738.5

COMPARATIVE STUDY OF FEATURE DETECTORS AND FILTERING METHODS IN IMAGE MATCHING

Andriy Fesiuk**, Yuriy Furgala
Ivan Franko National University of Lviv,
50 Drahomanova St., 79005 Lviv, Ukraine

Fesiuk A. V., Furgala Y. M. (2025). Comparative Study of Feature Detectors and Keypoint Filters in Image Matching. *Electronics and Information Technologies*, 31, 71–88. https://doi.org/10.30970/eli.31.7

ABSTRACT

Background. In modern computer vision, the accuracy and reliability of image matching primarily depend on the quality of local feature processing. False correspondences, which arise from changes in scale, illumination, or repetitive structures, have the potential to distort a scene's geometric model. Therefore, applying filtering algorithms that can distinguish informative matches from noise becomes an important step. Despite significant progress, most research focuses only on specific combinations of detectors and filtering methods, which prevents a comprehensive understanding of their interaction.

Materials and Methods. To investigate this issue, a series of experiments was conducted, and a representative subset from the Photo Tourism dataset was selected. Keypoint detection and description were performed, followed by matching, outlier filtering, and quantitative evaluation. The comparison involved the SIFT, SURF, KAZE, AKAZE, ORB, and BRISK detectors combined with the RANSAC, LMedS, RHO, GMS, VFC, and LPM filtering methods. For the evaluation, metrics such as the Fisher Criterion, IQR Separability, Whisker Gap, and a custom-developed metric called SMS were applied.

Results and Discussion. The investigation revealed that performance varies significantly among the detectors: binary descriptors offer much higher processing speeds. In contrast, methods using floating-point descriptors are more informative but require more computational resources. The hierarchy of filtering methods was consistent across all setups: VFC achieved the highest quality based on separability metrics, while LPM showed the most considerable difference between the distribution boundaries. RANSAC and LMedS remain classic benchmarks, while GMS and RHO serve as fast, compromise alternatives.

Conclusion. The results show that image matching effectiveness depends on the combination of the detector, the number of keypoints, and the filtering method. A comprehensive approach enables the selection of the right strategies for specific tasks, ranging from applications that require fast processing to scenarios that necessitate maximum separability or boundary error control. The analysis and metrics used provide a basis for future research and improvements in practical computer vision systems.

Keywords: feature detection, keypoint descriptors, image matching, match filtering.

INTRODUCTION

In the modern world, the amount of visual data is increasing rapidly - from countless photos on social media to continuous video streams from surveillance cameras and autonomous systems. For automated analysis of this large flow of information, local feature processing algorithms are essential. Detecting keypoints is fundamental for various computer vision tasks, including three-dimensional reconstruction from images or Structure

from Motion, panorama creation, Simultaneous Localization and Mapping (SLAM), and visual odometry [1-5].

A typical image processing pipeline involves detecting keypoints, describing them, and matching based on the similarity of these descriptors. However, the initial set of correspondences often contains many false matches or outliers, which are caused by visual ambiguity, repetitive textures, or lighting changes. Even a few outliers can significantly affect the estimation of geometric models. Therefore, filtering is an important step that eliminates false matches and returns a set of pairs that align geometrically.

To reduce the number of outliers, a wide range of methods has been proposed - from classic parametric techniques, notably RANSAC (Random Sample Consensus) [6] and its variations LMedS (Least-Median of Squares) [7] and RHO (Randomized Hough Transform) [8], to non-parametric approaches such as GMS (Grid-based Motion Statistics) [9], VFC (Vector Field Consensus) [10], and LPM (Locality Preserving Matching) [11]. Previous studies have primarily focused on specific aspects of the problem, either by comparing detectors and descriptors [12-14] or by analyzing the performance of filters in limited combinations [15-17]. The main contribution of this work lies in evaluating the interaction between detection and filtering methods.

A combination of six detectors - SIFT (Scale-Invariant Feature Transform) [18], SURF (Speeded-Up Robust Features) [19], KAZE [20], AKAZE (Accelerated KAZE) [21], ORB (Oriented FAST and Rotated BRIEF) [22], and BRISK (Binary Robust Invariant Scalable Keypoints) [23] - with six filtering methods - RANSAC, LMedS, RHO, GMS, VFC, and LPM - was examined. It is demonstrated how specific detector-filter pairings affect execution time and the overall matching quality. The experiments were performed on a subset of the public Image Matching Challenge Photo Tourism dataset [24], which features scenes with significant variations in viewpoints and scales of various architectural landmarks. Unlike traditional metrics such as precision and recall, our primary focus was on the ability of the filters to separate, specifically to increase the statistical distance between the distributions of similar and dissimilar images. To achieve this, we applied a set of specialized metrics: the Fisher Criterion, IQR (Interquartile Range) Separability, Whisker Gap, and a proposed metric, Sigma-Margin Separability (SMS). Together, these metrics offer a comprehensive view of the distributions' central parts and their tail regions.

MATERIALS AND METHODS

A software pipeline was developed and implemented to compare filtering methods. It includes several sequential stages: dataset preparation, keypoint detection and description, initial matching, outlier filtering, and an evaluation of effectiveness using a set of class separability metrics.

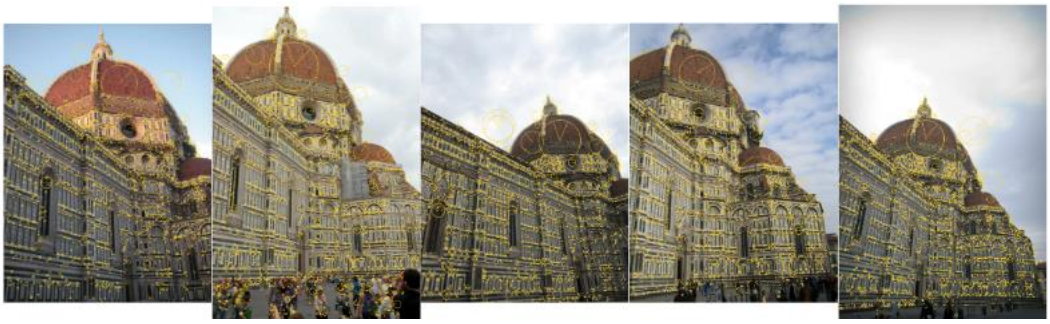
The public Image Matching Challenge Photo Tourism dataset, which includes architectural scenes with a wide range of viewpoints and lighting conditions, was used as the data source. Since the full dataset is resource-intensive and contains a large amount of data, this study used an automated process to select representative images for each scene. Initially, local features were detected and described for each frame using the SIFT method, which is known for its robustness to scale and viewpoint changes and its ability to assess an image's informativeness. A normalized, averaged vector of descriptors then represented each image. From this, a set of candidate pairs was generated by identifying, for each frame, the nearest neighbors based on cosine similarity, thereby reducing the number of comparisons without losing meaningful intra-scene relationships. For each candidate pair, matching was performed using the Brute-Force method with Lowe's ratio test [18], followed by an evaluation of geometric consistency using RANSAC with a homography model [25]. Utilizing RANSAC at this stage is a key methodological choice; previous research [26] has shown that combining this method with SIFT yields the best results compared to other methods within the USAC family [27]. Based on these similarity

values, a symmetric similarity matrix was created, from which the densest k-subset with the highest average mutual similarity was selected, ensuring consistency among images and eliminating duplicates. As a result, 10 images from 10 classes were selected, forming the final test set of 100 frames. Examples of images selected into the dataset are presented with keypoints detected in **Fig. 1a** and with keypoints matched in **Fig. 1b**.

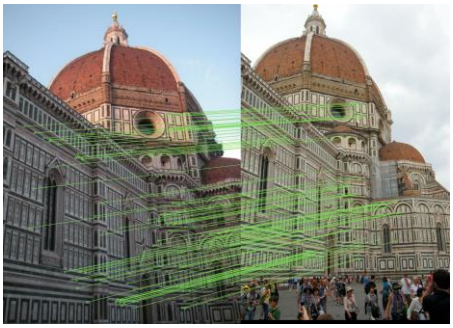
In this study, six methods for detecting keypoints were used: SIFT, SURF, ORB, BRISK, AKAZE, and KAZE. The initial number of keypoints for each detector was different. This is because only SIFT and ORB have direct parameters to limit the number of features, while SURF, BRISK, AKAZE, and KAZE are configured using a related threshold. The response attribute [28] was used to restrict and align the number of keypoints between detectors, and only the top features were selected for further analysis. This attribute, which depends on the detector's internal quality assessment of each point, is a standard practice for ensuring a fair comparison and serves as a universal criterion for sorting features; however, its physical meaning and mathematical basis are unique to each algorithm.

- SIFT: The amplitude of the extremum in the Difference of Gaussians (DoG) pyramid, which indicates contrast [18].
- SURF: The determinant of the Hessian matrix, utilized for identifying blob structures [19].
- ORB та BRISK: Metrics based on corner detectors - Harris for ORB [22] and AGAST for BRISK [23].
- KAZE та AKAZE: The determinant of the Hessian matrix calculated in a non-linear scale-space, which enables improved preservation of object boundaries compared to traditional Gaussian blurring.

In our study, the number of keypoints ranged from 500 to 5000, thereby enabling the assessment of the impact of keypoint density on subsequent processing stages.



(a)



(b)

Fig. 1. Example of images in dataset (Florence Cathedral): a) with keypoints detected (marked yellow); b) with keypoints matched (green lines).

The Brute Force matcher method was employed with a search for the two nearest neighbors (k = 2). The distance measure was chosen based on the descriptor type: Euclidean distance for floating-point descriptors and Hamming distance for binary descriptors [28]. After identifying candidate pairs, Lowe's ratio test was applied with a threshold of 0.75. The choice of this value was deliberate, as its effectiveness and appropriateness were evaluated in our earlier research [29]. This research demonstrated that this value provides a reliable balance between the number and quality of initial matches. This step removes ambiguous correspondences and minimizes the effect of random coincidences, thereby forming the initial set of matches for subsequent geometric verification and filtering. The matching time for each image pair was measured as the combined duration of the Brute-Force matching process and the subsequent Lowe's ratio test verification.

The primary focus of the study is a comparison of six filtering methods, representing three different strategies for outlier rejection. Homography was used as the base model for all methods requiring geometric verification. The investigation employed parametric methods, including the baseline RANSAC with its iterative approach, its variant, LMedS, which minimizes the median of errors, and the optimized RHO, which prioritizes sampling high-quality matches. Non-parametric methods were also examined, including VFC, which evaluates the smoothness of the match vector field, and LPM, which verifies the preservation of local neighborhood structure. Additionally, the statistical method GMS was analyzed, utilizing a grid to assess motion consistency in local regions. This approach enables an investigation into how filters based on a geometric model compare to methods that use spatial or statistical approaches, without relying on model parameterization.

The experiments were conducted on a 2019 MacBook Pro with an Intel Core i9-9880H processor, clocked at 2.30 GHz, featuring eight physical cores/16 threads, and 32 GB of RAM, running macOS 15.6.1. The software versions used were Python 3.9.13 and OpenCV 4.10.0. Parallel processing was achieved using a process pool with eight workers; OpenCV's internal parallelism was disabled to reduce variability [28]. The detection, matching, and filtering were measured within their respective procedures.

To quantitatively assess the effectiveness of each filtering method, a methodology was established based on analyzing class separability. Central to this approach is the similarity coefficient S, which is calculated for each pair of images (i, j) after the filtering process has been applied. This coefficient is a normalized measure that indicates the proportion of "good" matches relative to the total number of features available.

$$S_{i,j} = 100 \cdot \frac{N_{good}}{\min(N_i, N_j)}, \quad (1)$$

where N_{good} is the number of matches remaining after filtering, and N_i and N_j are the total number of keypoints in images i and j, respectively. Based on this coefficient, all image pairs were divided into two categories: "Similar" for images that are similar, and "Different" for images that are different.

The effectiveness of a filtering method is determined by its ability to create two well-separated distributions of the coefficient S for these two classes. After processing scenes with an ideal filter, the S values should be high for similar images and low for dissimilar ones. A comprehensive set of four metrics was used to evaluate this separability quantitatively.

- Fisher Criterion: A classic statistical measure that compares the distance between the centers of the distributions (between-class variance) to their internal consistency (within-class variance). A higher value indicates better separation.

$$F = \frac{(\mu_{sim} - \mu_{diff})^2}{\sigma_{sim}^2 + \sigma_{diff}^2}, \quad (2)$$

where μ та σ^2 are the mean and variance of the respective distributions.

- IQR Separability Coefficient: An outlier-robust metric that quantitatively measures the gap between the central 50% of data in each distribution (the interquartile ranges), directly showing the visual separation of the "boxes" in box plots.

$$IQR\ Separability = \frac{Q1_{sim} - Q3_{diff}}{Q3_{sim} - Q1_{diff}}, \quad (3)$$

where Q1 and Q3 are the first and third quartiles of the respective distributions.

- Whisker Gap: A straightforward numerical estimate of the distance between the outermost boundaries of the main data ranges, which reflects the visual gap between the "whiskers" on box plots calculated using the standard Tukey's method with a coefficient of $1.5 \cdot IQR$.

$$Whisker\ Gap = Q1_{sim} - Q4_{diff} \quad (4)$$

where $Q1_{sim}$ is the lower bound of the similar images distribution, and $Q4_{diff}$ is the upper bound of the dissimilar images distribution.

- Sigma-Margin Separability (SMS) Coefficient: To thoroughly assess class separability, we have developed a new metric called Sigma-Margin Separability, in addition to traditional methods. The motivation behind this development comes from the limitations of existing metrics. For instance, the Fisher Criterion does not directly verify whether a "guaranteed margin" exists between the distributions, particularly in their tails. Similarly, metrics like IQR Separability are robust to outliers but achieve this by focusing only on the central 50% of the data, completely ignoring behavior at the distribution boundaries. While the Whisker Gap does consider the outer ranges, it is based on a structural definition rather than a probabilistic one. In our case, margins between the statistical boundaries of the "Similar" and "Different" distributions were evaluated. These boundaries are determined by the *k*-sigma rule ($\mu \pm k \cdot \sigma$). The metric is calculated using the formula:

$$SMS(k) = \frac{(\mu_{sim} - k \cdot \sigma_{sim}) - (\mu_{diff} + k \cdot \sigma_{diff})}{\mu_{sim} - \mu_{diff}} \quad (5)$$

where the parameter *k* controls the stringency of the criterion, in this work, values of *k* = 1, 2, and 3 were used, which correspond to checking the margin between boundaries encompassing approximately 68%, 95%, and 99.7% of the data in each distribution, respectively. A larger value of *k* indicates a more rigorous evaluation. The numerator in the formula represents the "k-sigma margin" – the distance between the lower bound of the "Similar" distribution and the upper bound of the "Different" distribution. The denominator serves as a normalization factor. A positive SMS value indicates high separation reliability, showing a "guaranteed margin" between the k-sigma boundaries. Conversely, a negative value suggests overlap between them. A more detailed statistical analysis of SMS, including sensitivity to distributional shifts, sample-size effects, and robustness, is beyond the scope of this paper and will be covered in a separate study.

Thus, using this set of metrics enables a thorough evaluation of filter effectiveness by examining different aspects of separability, ranging from traditional statistical methods to reliability assessments with strict criteria and intuitively understandable indicators that visually demonstrate the overlap of distributions. For all the listed metrics, higher values denote better class separation. This approach offers a more complete understanding of how detection methods perform when combined with various filtering techniques.

RESULTS AND DISCUSSION

This section provides a comprehensive analysis of the research findings. Initially, the computational performance of the methods is evaluated across three essential data processing stages: detection and description, matching, and filtering. Subsequently, a detailed assessment of the filtering quality is conducted for each detector.

The results presented in **Table 1** show the average number of detected keypoints, the time needed for their detection and description, and the normalized efficiency, calculated as the average time for detection and description per 1000 keypoints. Normalizing these metrics enables an objective comparison of the detectors' efficiency by mitigating the influence of the varying number of keypoints they generate.

The obtained results indicate a distinct performance benefit for the binary methods. While each method produced a different number of keypoints with the given input parameters, the ORB method remains the clear leader in both total processing time and normalized efficiency. BRISK takes second place and significantly outperforms methods that use floating-point descriptors.

In contrast, methods such as SIFT, SURF, and KAZE, which are based on floating-point descriptors, were more resource-intensive. KAZE shows the lowest performance, with its normalized time approximately 3.8 times higher than SIFT's and nearly 2.5 times higher than SURF's; this aligns with the computational complexity of the non-linear diffusion scale-space on which the method is based. The accelerated version, AKAZE, operates much faster; however, in terms of normalized efficiency, it still significantly lags behind the other binary methods.

The dependency of the average matching time on the number of keypoints for a single image pair is shown in **Fig. 2**. The slowest method was SIFT, which is over 2.5 times slower than ORB. Analyzing scalability - the rate at which time increases with more points - shows that KAZE is the most robust to an increasing number of features, with the lowest rate of about 90 ms for each additional 1000 points, nearly matching ORB. In contrast, SIFT scales much worse, exceeding KAZE's rate by more than three times. Other methods vary in performance. SURF offers moderate speed and scalability. BRISK and AKAZE work well with a few keypoints but become less efficient as the number increases, unlike ORB and KAZE.

The filtering time is shown in **Fig. 3**, where a performance hierarchy of the investigated methods is visible, which was consistently maintained for all six detection methods. This

Table 1. **Time characteristics and computational efficiency of keypoint detection and description.**

Method	Avg number of keypoints	Avg time of detection, ms	Avg time of description, ms	Avg total time, ms	Avg total time per 1000 keypoints, ms
SIFT	6811	213.37	358.34	571.58	85.1
SURF	8486	273.37	834.56	1108.03	131.81
ORB	8397	37.46	16.17	53.63	6.38
BRISK	9517	98.47	77.26	175.73	18.57
AKAZE	8796	197.4	260.92	458.31	52.88
KAZE	8404	1286.61	1325.31	2611.92	318.49

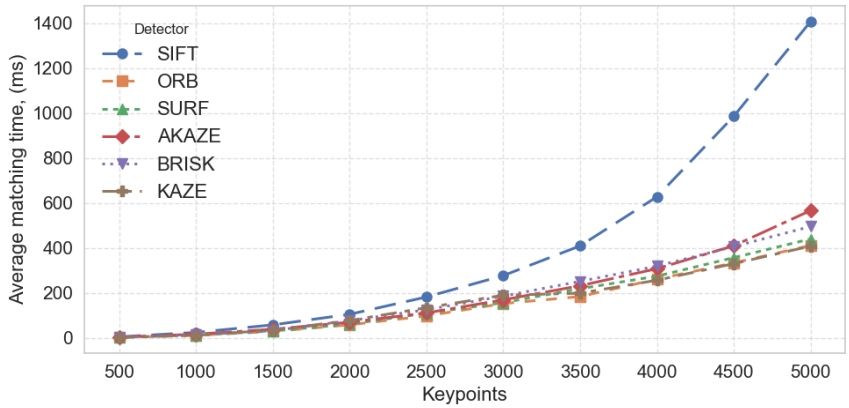


Fig. 2. Average matching time per image pair vs. number of keypoints for different feature detectors.

stability suggests that the choice of filtering method has a significant impact on the matching process. RHO and VFC demonstrated the highest speed and scalability, with an increase in the number of keypoints having almost no effect on processing time. In contrast, classic RANSAC proved to be the least effective, lagging behind the leaders in speed by an order of magnitude and exhibiting increased processing time as the number of points increased. GMS, LMedS, and LPM occupied an intermediate position, with GMS typically being the fastest and LPM the slowest.

The type of detector affects the time values. However, the relative ranking of the filtering methods stays the same. For the SIFT, SURF, and AKAZE detectors, the performance difference between the high-speed RHO and VFC and the slower RANSAC is most noticeable. For the ORB, BRISK, and KAZE detectors, the overall execution times are lower.

An analysis of the results for the SIFT detector, shown in **Fig. 4**, indicates that increasing the keypoint limit consistently improves all filtering quality metrics. The effectiveness plateaus at around 3000-4000 keypoints, where adding more features no longer provides a significant benefit.

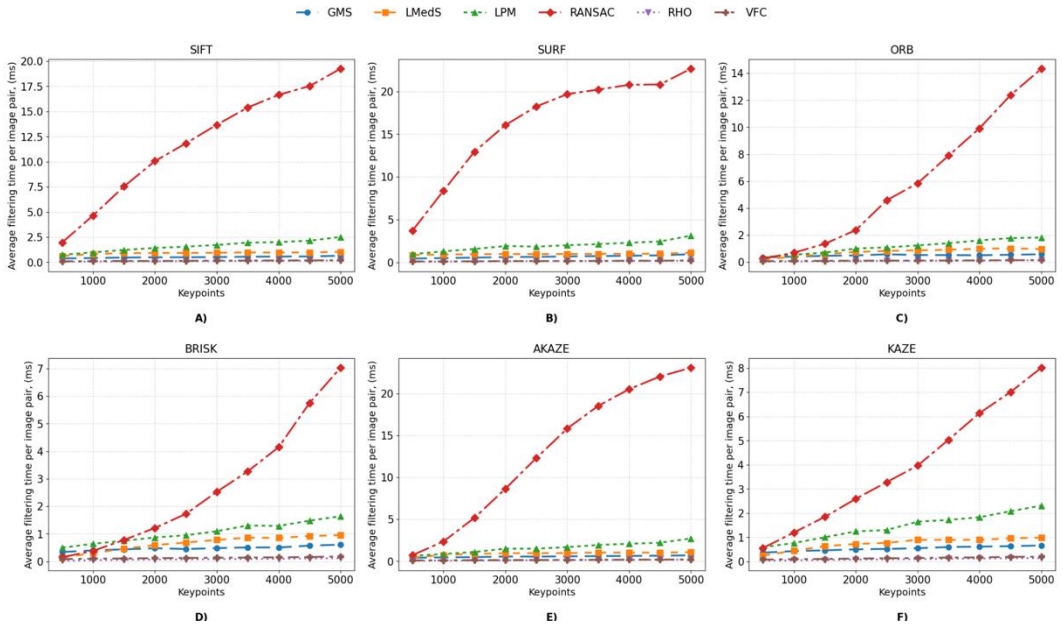


Fig. 3. Average filtering time per image pair vs. number of keypoints for different feature detectors: A) SIFT; B) SURF; C) ORB; D) BRISK; E) AKAZE; F) KAZE.

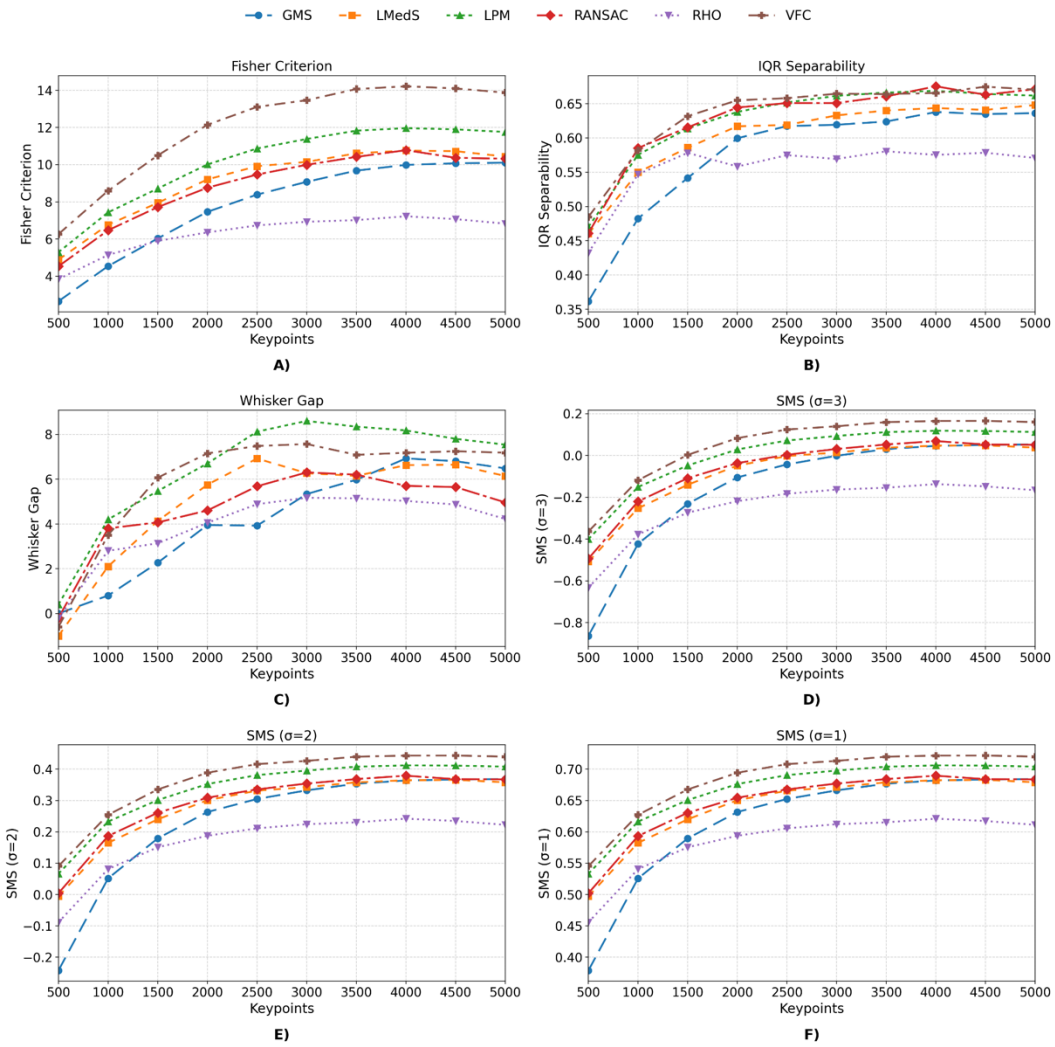


Fig. 4. Filtering quality metrics for SIFT: A) Fisher Criterion; B) IQR Separability; C) Whisker Gap; and the SMS coefficients for D) σ=1, E) σ=2, and F) σ=3

According to the Fisher Criterion, the results of the VFC method increase by approximately 2.3 times when moving from 500 to 4000 keypoints. In the saturation zone at 4000 keypoints, it exceeds LPM by nearly 20% and RANSAC and LMedS by about 30-35%. This shows that VFC achieves the best balance between between-class separation and within-class variance, while GMS, with 40% lower metrics, and RHO, which is almost twice as poor, perform worse.

For the Whisker Gap metric, the best results are achieved by LPM. Starting from the 2500 keypoint range, it creates the most significant gap between the classes, surpassing VFC by approximately 10-15%. This makes LPM a good choice where it is important to minimize errors at the extreme boundaries of the data, which are often associated with outliers. The GMS method also exhibits a rapid increase in values after 2500 keypoints, and in the 4000-5000 range, it yields higher values than the RANSAC, LMedS, and RHO methods.

In contrast, for the IQR Separability metric at 4000 keypoints, the classic RANSAC shows a slight edge over VFC. This suggests that RANSAC works well with reliable matches but less effectively on more ambiguous data. The values for LMedS are similar to RANSAC's, while GMS is consistently 5-10% lower than VFC, and for RHO, this gap widens to 10-15%.

The values of the SMS coefficients demonstrate the superiority of VFC. In the 3000-4000 keypoint zone for σ = 1, the VFC method outperforms LPM by approximately 2-3% and RANSAC by 5-6%. For σ = 2, the advantage of VFC over LPM increases to approximately 40%. Furthermore, the σ = 3 values for VFC are in the positive region starting from 1500 keypoints, which indicates reliable separation even in the tail regions of the distributions. GMS is roughly 10-15% less effective than VFC across the entire coefficient range, while RHO has the lowest values, indicating sensitivity at the extreme points.

Unlike SIFT, the SURF detector exhibits a distinct performance trend, as illustrated in **Fig. 5**. All metrics display a sharp rise until reaching a peak between 1000 and 2000 keypoints, followed by a gradual decline. This pattern aligns with the presumption that at higher keypoint counts, more weak correspondences enter the SURF sample, which reduces the separation between the classes.

According to the Fisher Criterion, the best result is shown by VFC. Its metric reaches a peak at 1500 keypoints, after which it decreases by approximately 20%, while remaining superior to the other methods. At its maximum, VFC surpasses LPM by about 10%, and RANSAC and LMedS by nearly 20% and 35%, respectively. The values for RHO from 500

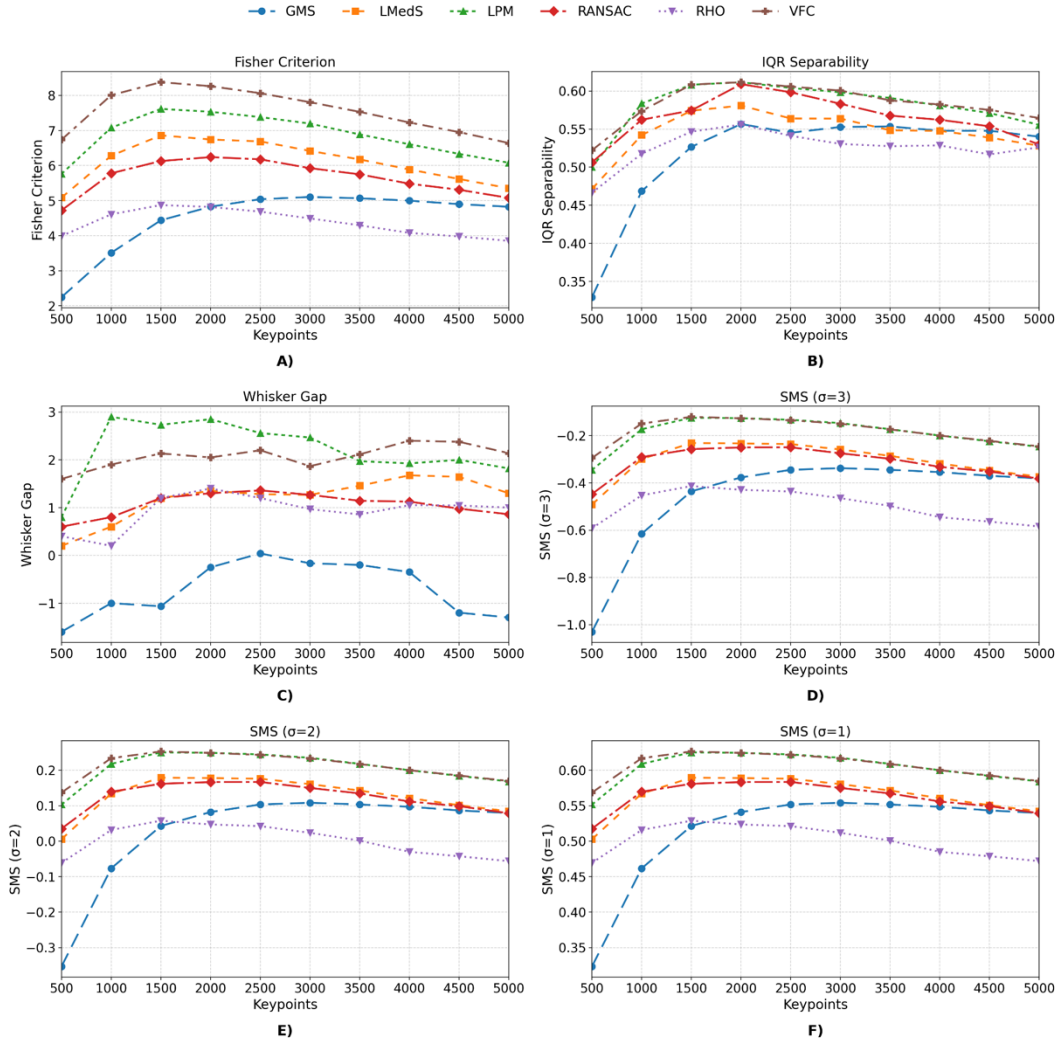


Fig. 5. Filtering quality metrics for SURF: A) Fisher Criterion; B) IQR Separability; C) Whisker Gap; and the SMS coefficients for D) σ=1, E) σ=2, and F) σ=3

to 2000 keypoints are higher than those of GMS; however, after 2500 keypoints, the metrics for GMS increase noticeably, and at 5000 keypoints, they are close to those of RANSAC. Despite this, the results for GMS and RHO are significantly lower than those of the other methods. A similar trend is observed for IQR Separability. The maximum also occurs in the 1500-2000 keypoint range, where RANSAC is almost on par with VFC; however, as the number of keypoints increases, its performance declines more rapidly. At more than 3500 keypoints, the GMS method partially overtakes the LMedS method, indicating that the method's efficiency improves with an increase in features.

For the Whisker Gap metric, the ranking of the methods changes. Within the optimal range of 1000-3000 keypoints, LPM shows the largest gap, exceeding VFC by 25-40% and RANSAC by over 1.3 times. However, as the number of keypoints increases further, the advantage shifts to VFC, which demonstrates more stable behavior. The GMS method performs the worst, with its values consistently negative throughout the entire range, indicating an overlap of distribution tails. However, this trend reverses at 2500 keypoints, where the metrics briefly turn positive before gradually decreasing.

The complex SMS coefficients confirm the same pattern, with peak efficiency at 1000-2000 keypoints. VFC leads in all three metrics, especially for $\sigma = 3$, where it exhibits the least negative values, indicating greater robustness to anomalous correspondence pairs. This highlights SURF's sensitivity to oversaturation with weak correspondences at large ranges and the importance of using a controlled keypoint limit in practical applications.

The results for the KAZE detector, shown in **Fig. 6**, demonstrate that increasing the number of keypoints consistently improves filtering quality. Efficiency levels off at 3500-5000 points. In contrast to SURF, the KAZE detector does not exhibit a decline in values at higher keypoint ranges, indicating a higher stability of its features. Although KAZE is weaker than SIFT in absolute metric scores, the performance hierarchy of the filters remains similar.

Based on the Fisher Criterion, metric values increase and plateau around 4000-5000 points. In this range, VFC shows the best performance, with its metric approximately 7% higher than LPM's and 25% higher than RANSAC's, demonstrating superiority over GMS of roughly 1.5 times. Compared to RHO, VFC's improvement is nearly twice as great. This suggests that the KAZE and VFC setup offers the most balanced performance for maximizing class separation with constant variance. A similar trend is observed with IQR Separability, where the central quartile separation stabilizes at approximately 3000 points, again favoring VFC. Meanwhile, LPM, RANSAC, LMedS, and GMS methods fall behind by a few percentage points, and RHO exhibits much lower effectiveness.

For the Whisker Gap metric, LPM leads in the 500-1500 keypoint range. However, from 2000 to 5000 points, VFC achieves the highest values, approximately 15% higher. Meanwhile, LPM significantly outperforms RANSAC and LMedS, for instance, at 3000 keypoints, by factors of approximately 1.4 and 1.7, respectively. This makes LPM the preferred choice for tasks where reducing the overlap of boundary values is essential.

The complex SMS coefficients consistently demonstrate VFC's leadership. It exceeds its competitors in all three metrics, especially at $\sigma = 3$, where its values are 20–30% closer to zero than those of RANSAC and LMedS. This suggests VFC offers the optimal balance between separation and robustness against anomalies. Conversely, RHO shows the worst values, being the furthest from zero.

The results for the binary detector ORB, presented in **Fig. 7**, show that increasing the keypoint limit consistently improves all filtering quality metrics, reaching a saturation point at 350-5000 keypoints. In terms of curve characteristics, ORB is similar to KAZE, as its efficiency increases with more data, unlike SURF, where a decline occurs after an early peak. In absolute metric values, ORB is expectedly inferior to SIFT.

According to the Fisher Criterion, VFC provides the best values across the entire range. In the zone of stable efficiency, its values are about 5-7% higher than those of LPM, nearly 7% higher than RANSAC and LMedS, and 25-30% higher than GMS. Compared to

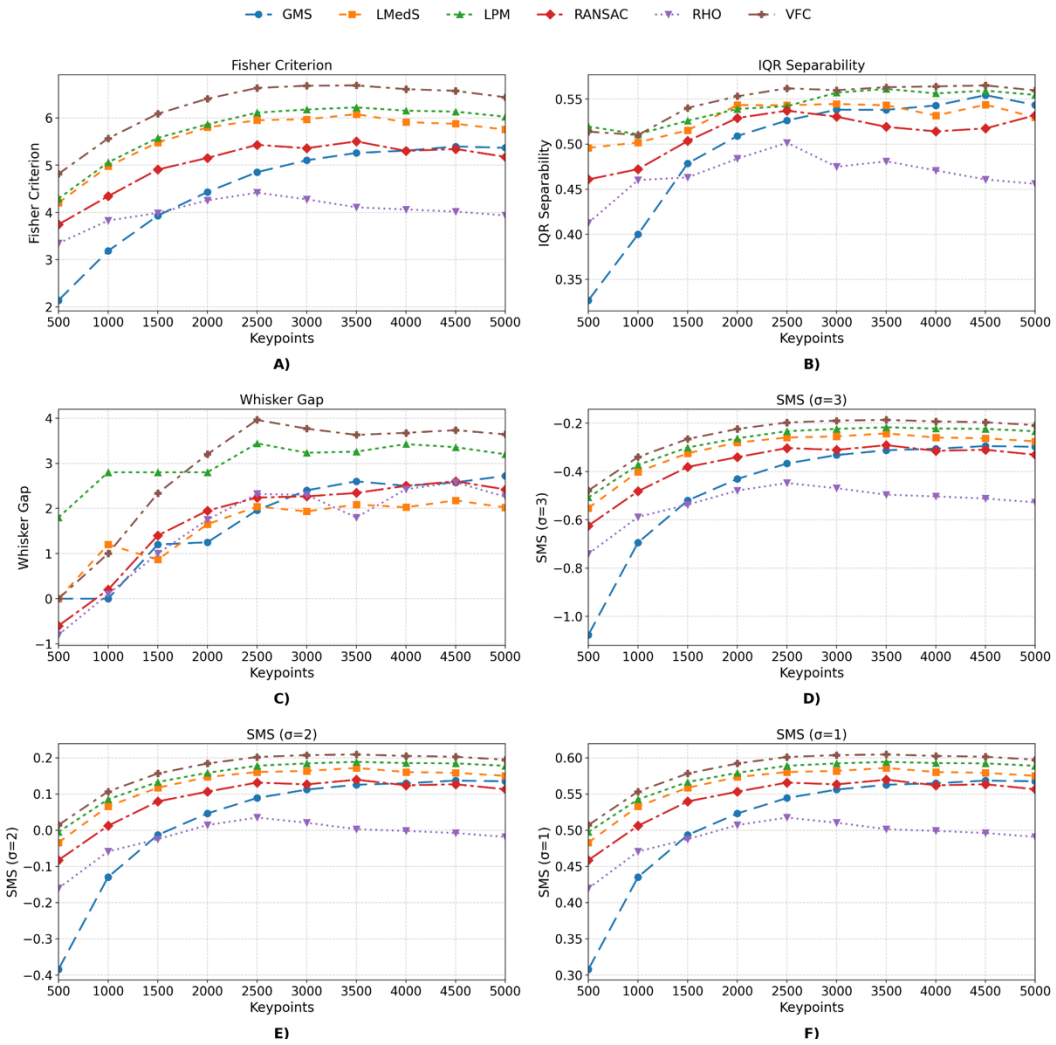


Fig. 6. Filtering quality metrics for KAZE: A) Fisher Criterion; B) IQR Separability; C) Whisker Gap; and the SMS coefficients for D) σ=1, E) σ=2, and F) σ=3.

RHO, its advantage exceeds 1.5 times. This indicates that VFC most effectively converts additional keypoints into stable inliers, maintaining a favorable balance between class separation and within-class variance.

The separation of the central quartiles increases with the number of keypoints and stabilizes around 3500 points. In the range of 3500 to 5000 keypoints, VFC outperforms LPM by 2-6%, and RANSAC and LMedS by approximately 8-10%. The GMS method is, on average, 7-14% lower than VFC, and RHO is more than 20% lower. Therefore, when the goal is to maximize the separation of the distribution centers, VFC maintains a consistent advantage, while RANSAC and LMedS serve as alternatives.

For the Whisker Gap metric, a partial overlap of the boundaries occurs across all filtering methods at low keypoint counts from 500 to 1000. Starting around 1500–2000 keypoints, the gap consistently turns positive and grows larger. With more keypoints, LPM and VFC produce the highest values; their gaps are usually 30–60% larger than those of RANSAC and LMedS. GMS and RHO show the smallest gaps, often at least 1.5 to 2 times smaller than the leaders. Practically, this suggests that when boundary value control is vital, it is best to use LPM or VFC with more than 2000 keypoints but avoid LPM with tiny samples.

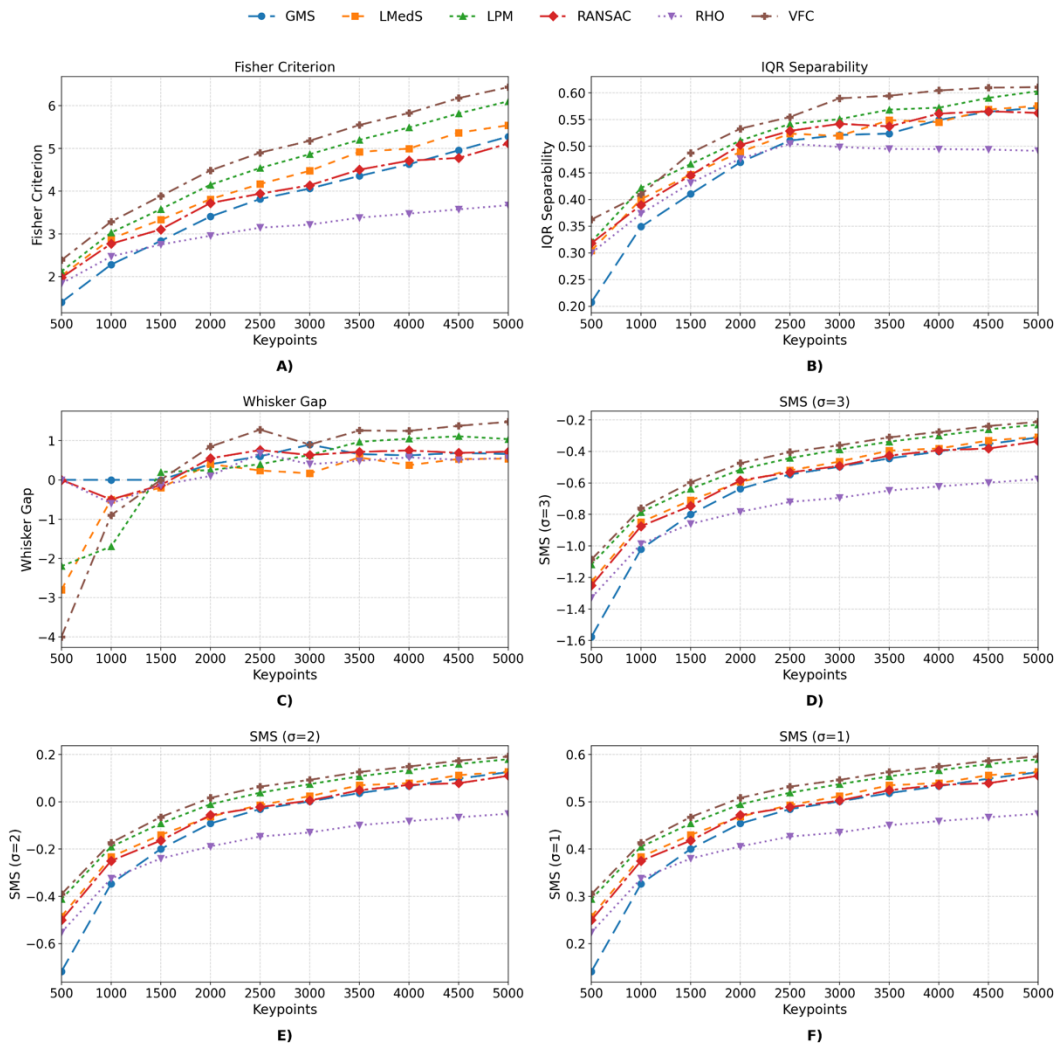


Fig. 7. Filtering quality metrics for ORB: A) Fisher Criterion; B) IQR Separability; C) Whisker Gap; and the SMS coefficients for D) σ=1, E) σ=2, and F) σ=3.

The above findings also apply to all SMS coefficient values; the curves reach saturation between 3,500 and 5,000 keypoints. VFC consistently exhibits the highest values across all SMS variants. For σ = 1, its advantage over LPM is modest, about 1-2%, but it surpasses RANSAC and LMedS by approximately 5-10%. When σ = 2, VFC's values turn positive at approximately 1500 keypoints, reaching a peak. For σ = 3, VFC remains 20-30% closer to zero than other methods, with RHO consistently demonstrating the worst results.

The results for the binary detector BRISK, presented in **Fig. 8**, show that increasing the keypoint limit consistently enhances all filtering quality metrics, which stabilize within the range of 3500-5000 points. Due to the nature of its curves, BRISK resembles KAZE and ORB, as the accumulation of valid correspondences gradually improves class separation without degradation, unlike SURF. In terms of absolute metric levels, BRISK is expectedly lower than SIFT, but at high numbers of points, it achieves a quality comparable to KAZE.

According to the Fisher Criterion, VFC delivers the best results across the entire range. In the saturation zone, its metric is about 14% higher than that of LPM, and nearly 17% and 30% higher than those of LMedS and RANSAC, respectively. Regarding GMS, its values at 5000 keypoints are similar to those of RANSAC and more than 1.5 times lower than

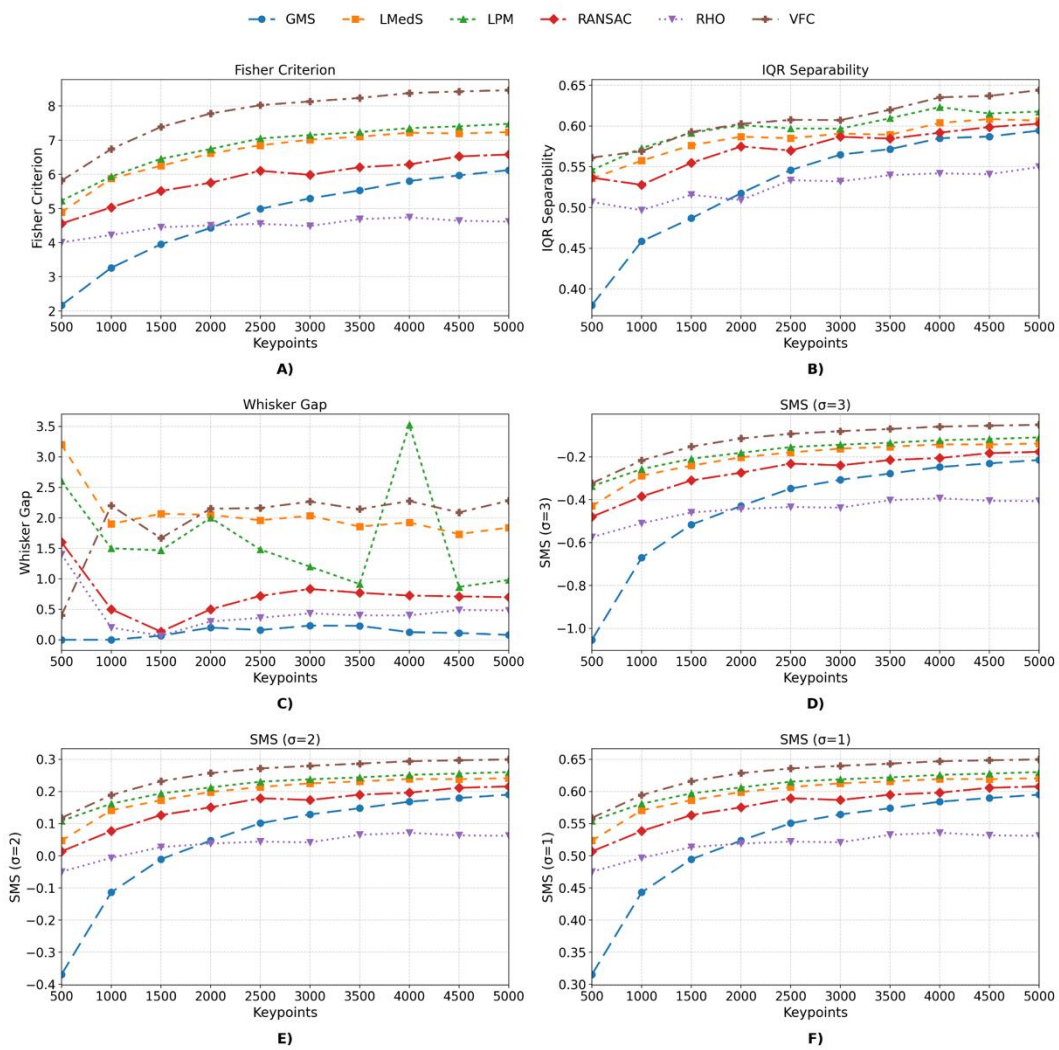


Fig. 8. Filtering quality metrics for BRISK: A) Fisher Criterion; B) IQR Separability; C) Whisker Gap; and the SMS coefficients for D) σ=1, E) σ=2, and F) σ=3.

those of VFC. This indicates that VFC most effectively converts additional data into stable correspondences, ensuring the optimal balance between class separation and within-class variance. A similar pattern is observed for IQR Separability, where VFC also shows the best results; in contrast, RHO performs the worst, outperforming GMS only in the range from 500 to 1500 keypoints.

For the Whisker Gap metric, the LMedS method starts strongest at the beginning of the range, then mostly maintains second place. Later, in the 1500-3000 keypoint range, VFC takes the top spot, with results about four times better than those of RANSAC. At 4000 keypoints, a noticeable peak occurs for LPM, where its results are roughly 35% higher than those of VFC. The RHO values are a few percent above GMS's, but their values are nearly four times lower than those of VFC and LMedS.

The complex SMS coefficients consistently confirm VFC's leadership across all metrics. Its advantage over LPM for σ=1 is approximately 3-4%, and it is about 5% higher than LMedS. For σ = 2 values, VFC's metrics are roughly twice as good as RANSAC's. According to the σ = 3 metric, VFC's results are closer to zero than those of the other methods. The RHO method has the lowest values for all SMS variations, except in the

interval between 500 and 2000 keypoints, where it outperforms GMS. Additionally, the RHO results are the most consistent across the entire range of keypoints.

The results for the AKAZE detector, as shown in **Fig. 9**, exhibit a different pattern than those for KAZE. Efficiency quickly rises at the start, peaks between 1500 and 2500 keypoints, and then generally stabilizes or declines slowly. This early peak behavior is more similar to SURF than to KAZE or ORB.

According to the Fisher Criterion, VFC yields the best result. Its metric peaks at 2000 keypoints, then decreases slightly by about 14% but remains the leader. At its peak, VFC exceeds LPM by 6% and LMedS by approximately 12%, and is almost twice as good as RHO. A different pattern was observed for the GMS method; after 3500 keypoints, its results level off, and the values are somewhat higher than those of RANSAC.

For the IQR Separability metric, the highest values are observed at 1500-2000 keypoints, where VFC has a slight advantage over LPM of 2.5%, although after 4000 keypoints, LPM takes the lead. LMedS outperforms RANSAC by 2-4%. GMS exhibits an interesting trend: its values are the lowest from 500 to 1500 points; at 2500 points, it surpasses RANSAC; and from 3500 keypoints to the end of the range, it competes with the

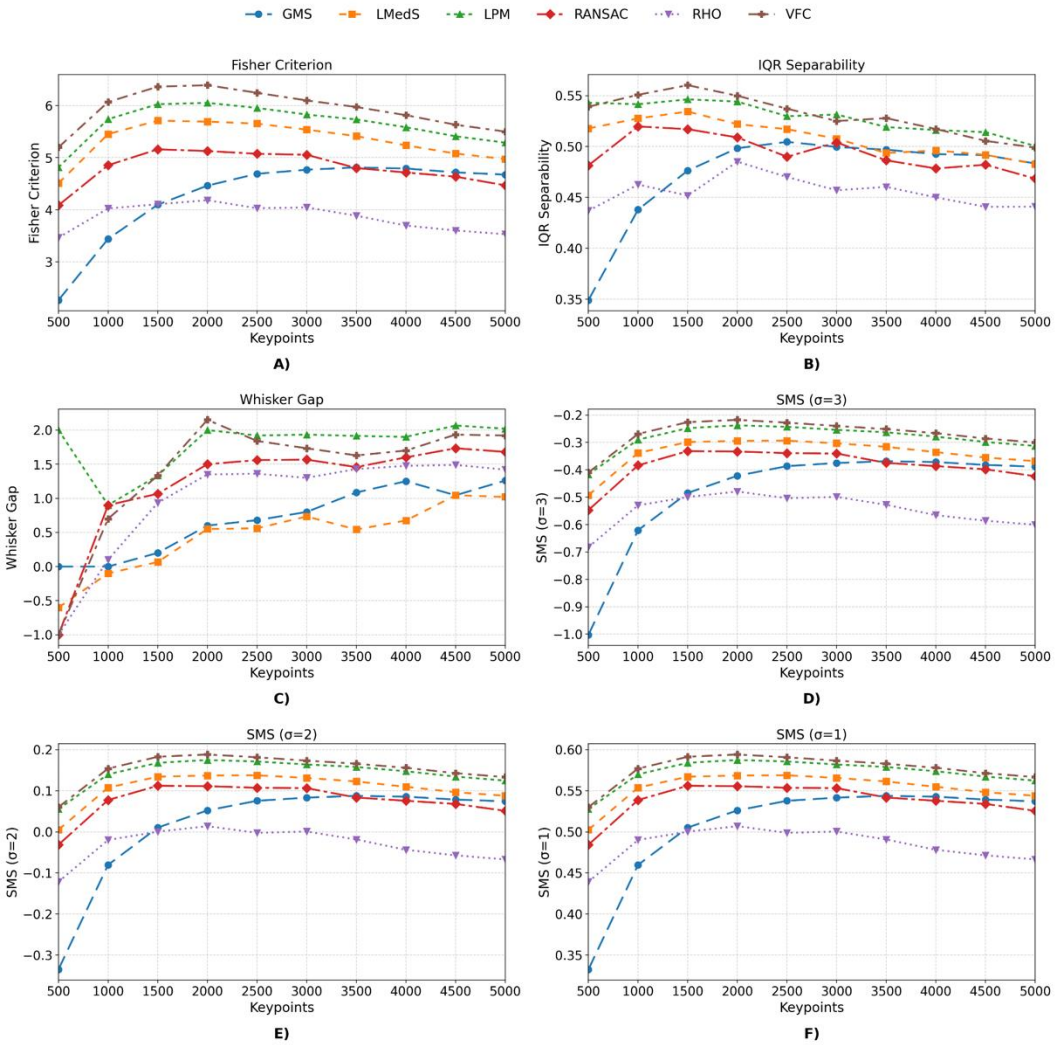


Fig. 9. Filtering quality metrics for AKAZE: A) Fisher Criterion; B) IQR Separability; C) Whisker Gap; and the SMS coefficients for D) σ=1, E) σ=2, and F) σ=3.

LMedS method. The RHO method shows the lowest values and tends to degrade after 2500 keypoints.

For the Whisker Gap metric, the ranking of methods shifts. Within the optimal range of 3000-5000 points, LPM exhibits the most significant gap, outperforming VFC by 5-20% and RANSAC by more than 1.3 times. The LMedS and GMS methods yield the worst results; for LMedS, this is particularly evident at 2500-3000 keypoints, where a significant gap is observed compared to other methods. In contrast, RHO shows values close to RANSAC's after 1500 points.

Across all SMS variations, VFC and LPM consistently rank first and second, with a slight difference of approximately 0.7-1%, peaking at 1500-2500 points. RHO performs the worst among the methods. The trend patterns for the AKAZE method resemble those seen with SURF, where GMS at over 1500 keypoints improves its metrics by nearly 1.5 times compared to RHO. At 4000-5000 keypoints, it surpasses RANSAC and approaches results similar to LMedS.

The results help in developing strategies for selecting the best configuration based on the specific task. For real-time applications, combining fast binary detectors such as ORB or BRISK with the VFC filter is effective, especially when limiting keypoints to the saturation zone for improved performance. LPM is recommended to reduce the overlap of distribution tails. If the goal is the highest matching quality, SIFT paired with VFC and a constrained keypoint limit is ideal to manage computational costs. When features are limited, SURF and AKAZE perform best near their efficiency peaks, offering an optimal quality-to-data ratio. Regardless of the detector used, RANSAC should be applied cautiously. Although it sometimes excels in the IQR Separability metric, its high computational demands and poor scalability often make it impractical for most modern systems.

CONCLUSION

This study presented a comparative analysis of various combinations of keypoint detection and filtering methods. Consistent patterns emerged: binary descriptors are characterized by high computational efficiency, whereas methods using floating-point descriptors tend to produce more informative but computationally intensive correspondences.

The analysis of filtering methods showed that VFC most often provides the most balanced separation of distributions and proves to be robust against noisy correspondences. At the same time, LPM is especially effective at controlling boundary cases by creating the greatest distance between the extreme boundaries. RANSAC and LMedS remain valuable as classic benchmarks, and GMS and RHO offer fast and lightweight alternatives. Despite differences between detectors, the relative ranking of the methods stays consistent, highlighting the universality of the identified patterns.

For tasks that require maximum separability, the best approach is to combine detectors with a stable saturation plateau, such as SIFT or KAZE, along with the VFC filter. When boundary error control is necessary, LPM is recommended. If processing speed is the primary concern, then combinations like ORB or BRISK with efficient filters, such as VFC or RHO, are suitable. The metrics used, including the proposed SMS coefficient, demonstrated their effectiveness in delivering a thorough assessment of filtering quality.

Thus, the research demonstrates that filtering correspondences in computer vision tasks cannot be considered independently of the choice of detector and keypoint parameters. A comprehensive approach allows for not only objective comparisons of methods but also the development of practical strategies to balance quality and computational costs. The main contribution of this work is to provide valuable recommendations for selecting optimal combinations for various applied scenarios. Future research could include integrating deep learning methods, employing other models to verify the geometric consistency of matches, and testing under more challenging scene conditions, particularly with varying scales, camera viewpoints, and changes in illumination.

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.

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] Soltanpour, S., & Joslin, E. (2025). A survey on feature-based and deep image stitching. *In Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Volume 3: VISAPP (pp. 777–788). https://doi.org/10.5220/0013368500003912

[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] Abaspur Kazerouni, I., Fitzgerald, L., Dooly, G., & Toal, D. (2022). A survey of state-of-the-art on visual SLAM. *Expert Systems with Applications, 205*, 117734. https://doi.org/10.1016/j.eswa.2022.117734

[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] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880. https://doi.org/10.1080/01621459.1984.10477105

[8] Chum, O., & Matas, J. (2005). Matching with PROSAC - Progressive sample consensus. *Proceedings of CVPR 2005* (pp. 220–226). https://doi.org/10.1109/CVPR.2005.221

[9] Bian, J.-W., Lin, W.-Y., Matsushita, Y., Yeung, S.-K., Nguyen, T.-D., & Cheng, M.-M. (2019). GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. *International Journal of Computer Vision, 128*(6), 1580–1593. https://doi.org/10.1007/s11263-019-01280-3

[10] Ma, J., Zhao, J., Tian, J., Yuille, A. L., & Tu, Z. (2014). Robust point matching via vector field consensus. *IEEE Transactions on Image Processing, 23*(4), 1706–1721. https://doi.org/10.1109/TIP.2014.2307478

[11] Ma, J., Zhao, J., Jiang, J., Zhou, H., & Guo, X. (2018). Locality Preserving Matching. *International Journal of Computer Vision, 127*(5), 512–531. https://doi.org/10.1007/s11263-018-1117-z

[12] Liao, Y., Di, Y., Zhu, K. *et al.* Local feature matching from detector-based to detector-free: a survey. *Appl Intell* 54, 3954–3989 (2024). https://doi.org/10.1007/s10489-024-05330-3
[13] Isik, M. (2024). Comprehensive empirical evaluation of feature extractors in computer vision. *PeerJ Computer Science, 10*, e2415.https://doi.org/10.7717/peerj-cs.2415
[14] S. A. Khan Tareen and R. H. Raza, "Potential of SIFT, SURF, KAZE, AKAZE, ORB, BRISK, AGAST, and 7 More Algorithms for Matching Extremely Variant Image Pairs," *2023 4th International Conference on Computing, Mathematics and Engineering Technologies,* pp. 1-6, 2023. https://doi.org/10.1109/iCoMET57998.2023.10099250
[15] Baráth, D., & Matas, J. (2022). Graph-Cut RANSAC: Local optimization on spatially coherent structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4961–4974. https://doi.org/10.1109/TPAMI.2021.3071812
[16] Yunge Cui, Yingming Hao, Qingxiao Wu, et al."An Optimized RANSAC for The Feature Matching of 3D LiDAR Point Cloud". *In Proceedings of the 2024 5th International Conference on Computing, Networks and Internet of Things (CNIOT '24). Association for Computing Machinery, New York, NY, USA*, pp. 287–291, 2024. https://doi.org/10.1145/3670105.3670153
[17] Rodríguez, M., Facciolo, G., & Morel, J.-M., "Robust Homography Estimation from Local Affine Maps". *Image Processing On Line*, 13, 65–89, 2023. https://doi.org/10.5201/ipol.2023.356
[18] 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
[19] 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
[20] 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
[21] 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
[22] 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
[23] 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
[24] Image Matching Challenge organizers. (2021). *Data — Image Matching Challenge 2021 (Phototourism subset)*. University of British Columbia. https://www.cs.ubc.ca/research/image-matching-challenge/2021/data
[25] Bazargani, H., Bilaniuk, O., & Laganière, R. (2018). A fast and robust homography scheme for real-time planar target detection. *Journal of Real-Time Image Processing, 15*(4), 739–758. https://doi.org/10.1007/s11554-015-0508-4
[26] 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
[27] M. Ivashechkin, D. Baráth, J. Matas, "USACv20: Robust Essential, Fundamental and Homography Matrix Estimation," 2021. https://doi.org/10.48550/arXiv.2104.05044
[28] 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.

[29] A. Fesiuk and Y. Furgala, “The Impact of Parameters on the Efficiency of Keypoints Detection and Description,” *2023 IEEE 13th International Conference on Electronics and Information Technologies (ELIT)*, Lviv, Ukraine, pp. 261-264, 2023. https://doi.org/10.1109/ELIT61488.2023.10310866

ПОРІВНЯЛЬНЕ ДОСЛІДЖЕННЯ ДЕТЕКТОРІВ ОСОБЛИВИХ ТОЧОК ТА МЕТОДІВ ФІЛЬТРАЦІЇ У ЗІСТАВЛЕННІ ЗОБРАЖЕНЬ

Андрій Фесюк**, Юрій Фургала
Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 Львів, Україна

АНОТАЦІЯ

Вступ. У сучасному комп’ютерному зорі точність і надійність зіставлення зображень значною мірою визначається якістю обробки локальних ознак. Хибні відповідності, що виникають через зміну масштабу, освітлення чи повторювані структури, здатні зруйнувати геометричну модель сцени. Тому важливим етапом стає застосування алгоритмів фільтрації, здатних відокремити інформативні збіги від шумових. Попри значний прогрес, більшість досліджень аналізують лише окремі комбінації детекторів та методів фільтрації, що не дозволяє сформувати цілісне уявлення про їхню взаємодію.

Матеріали та методи. Для дослідження цієї проблеми проведено експерименти, в межах яких виконано вибір репрезентативної підмножини з датасету Photo Tourism, виявлення й опис ключових точок, початкове зіставлення, фільтрацію викидів та кількісну оцінку. Порівняння проведено для детекторів SIFT, SURF, KAZE, AKAZE, ORB, BRISK у поєднанні з методами фільтрації RANSAC, LMedS, RHO, GMS, VFC, LPM. Для оцінки застосовано набір метрик, зокрема Fisher Criterion, IQR Separability, Whisker Gap та власну розробку SMS.

Результати. Дослідження показало, що продуктивність суттєво відрізняється між детекторами: бінарні дескриптори забезпечують значно вищу швидкодію, тоді як методи з дескрипторами з плаваючою комою демонструють кращу інформативність ціною більших витрат. Ієрархія методів фільтрації виявилася стабільною для всіх конфігурацій: найвищу якість за метриками роздільної здатності демонструє VFC, тоді як LPM забезпечує найбільший розрив між крайніми межами розподілів. RANSAC і LMedS залишаються класичними орієнтирами, а GMS і RHO є швидкими компромісними варіантами.

Висновки. Отримані результати показують, що ефективність зіставлення зображень визначається саме поєднанням детектора, кількості ключових точок та методу фільтрації. Комплексний підхід дозволяє обґрунтовано обирати стратегії під конкретні задачі: від застосунків, де вирішальною є швидкодія, до сценаріїв, де важливою є максимальна роздільна здатність чи контроль граничних помилок. Запропонований аналіз та використані метрики формують основу для подальших досліджень і вдосконалення практичних систем комп’ютерного зору.

Ключові слова: виявлення ознак, опис особливих точок, співпадіння, фільтрація.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
19 August, 2025	12 September, 2025	19 September , 2025	31 October, 2025