SIViP (2015) 9:1009–1019
DOI 10.1007/s11760-013-0535-1

ORIGINAL PAPER

Performance evaluation of keypoint detection and matching techniques on grayscale data

Hugo Proença

Received: 30 August 2012 / Revised: 31 May 2013 / Accepted: 15 July 2013 / Published online: 4 August 2013

Abstract The extraction of local photometric descriptors from images has been extensively reported in the computer vision literature. The main purpose of this paper is to provide an objective comparison between the performance of four of the most popular algorithms of this kind: SIFT, SURF, BRIEF and DAISY. Constraining our analysis to grayscale data, several major points distinguish this work from the previous evaluation initiatives: (1) A large amount of data were used, representing a broad range of *real-world* scenes; (2) an automated evaluation procedure was devised, in order to minimize subjectivity; and (3) we analyze the reliability of each algorithm not only in terms of the distances between corresponding feature descriptors but also of their order statistics. Also, the public availability of a new annotated data set is reported, which is suitable for the automated and statistically significant evaluation of keypoint detection and matching strategies.

Keywords Image registration · Feature extraction · Local descriptors · Interest points

1 Introduction

The relevance given to local descriptors has been increasing for different computer vision applications, including object recognition (e.g., [9,16,17]), image alignment [12], feature tracking [4] and robot navigation [6].

The effectiveness of such local descriptors directly corresponds to their invariance with respect to translation, scale, rotation, affine and perspective transforms. Various algorithms of this kind were proposed, based on intensity, color, texture or gradient information. In this work, we evaluate the performance of four of the most popular keypoint detection/matching techniques: SIFT [13], SURF [1], BRIEF [2] and DAISY [21]. Our analysis is constrained to grayscale data, which evidently corresponds to the most frequent scenario where these techniques are used. The justifications for the selected techniques are threefold: (1) the number of citations to these techniques in the computer vision literature; (2) the outstanding results reported by authors; and (3) the reputability of the information sources where these methods were published.

As given below, previous evaluation initiatives of this kind of techniques are reported in the literature. However, several drawbacks of these works can be enumerated: (1) Most used small data sets, which reduce the statistical significance of the results; (2) the used images regard specific scenarios, which makes generalization of the results difficult; and (3) subjective evaluation protocols were used, biasing the subsequent analysis.

Hence, the main purpose of this paper is to provide an objective comparison between the performance of keypoint detection/matching techniques. When compared to similar works, three distinguishing features can be enumerated: (1) A systematic and automated evaluation procedure was carried out, enabling to report objective results; (2) the used data comprise images acquired in *real-world* conditions and with jointly varying factors (changes in translation, scale, rotation, perspective, blur and lighting conditions); and (3) a large number of images were used in the evaluation, increasing the statistical significance of the results obtained.

As major conclusions, we found that—for all the algorithms evaluated—the reliability of keypoint correspondences varies most with respect to the order statistic [5]

H. Proença (✉)
Department of Computer Science, IT, Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal
e-mail: hugomcp@di.ubi.pt

of matching distances than with respect to the matching distances themselves, i.e., the relative distance between the descriptors of each correspondence reported is a better predictor of performance than the distance value in absolute terms. Also, each algorithm offers best results in specific ranges of the performance space, and too simplistic statements about their relative performance are erroneous.

1.1 Related performance evaluation initiatives

Schmid et al. [19] introduced two evaluation criteria for interest point detectors: the repeatability rate and the information content. Regarding the former, an improved version of the Harris detector was considered the best measure. The results for information content showed that the Harris detector outperformed the Heitger detector. Mikolajczyk and Schmid [14] evaluated local descriptors using recall/precision values, having considered changes in scale, rotations, blur and JPEG compression. In addition, they proposed an extension of SIFT, which outperformed all of the strategies tested in that study. Previously, the same authors compared the performance of SIFT, steerable filters, differential invariants, complex filters, moment invariants and cross-correlation for different types of interest points [15]. Focusing on the SIFT method, Tao et al. [20] surveyed research on SIFT-based object recognition and compared the performance of SIFT to its variants. They used recall/precision, repeatability rates and ROC curves as the main performance measures and concluded that SIFT behaves comparably to GLOH descriptors. Ke and Sukthankar [11] concluded that the PCA-SIFT variant is more distinctive, robust to image transforms and compact than the original SIFT approach. Carneiro and Jepson [3] compared the performance of their region descriptor, based on complex-valued steerable filters, to descriptors based on differential invariants. They concluded that their method leads to better performance under common illumination changes and 2-D rotation and obtains similar results in terms of scale changes. Randen and Husoy [18] compared the performance of different convolution kernels in local image patches, having observed that among the tested filters (Law, Gabor, FIR, DCT and wavelet decompositions), none consistently outranked the others. Rather, performance is strongly dependent on the application and filter parameterizations.

For color-based object recognition, Sande et al. [23] studied the invariance and distinctiveness of color descriptors with respect to photometric transformations. This study concluded that invariance to light intensity is possible to obtain, while changes in light color consistently affect performance. This study suggests that when choosing a single descriptor with no prior knowledge about the data set, object and scene categories, a variant of SIFT computed in the opponent color space is recommended. Preliminary results from the same team were given in [22].

1.2 Proposed performance evaluation protocol

A cohesive overview of the evaluation protocol devised for this work is depicted in Fig. 1:

1. We collected a data set of roughly planar scenes of indoor and outdoor environments. These images regard a broad range of scenes and are highly heterogeneous in terms of lighting conditions.
2. For every pair of images of the same scene (I_1 and I_2), a set of k_i point correspondences were manually annotated.
3. A homography was found per pair of images, representing the transformation of any position in I_1 to a position in I_2.
4. Keypoints were detected and matched for every pair of images, according to the four techniques considered.
5. The goodness for each pair of matched keypoints was measured, with respect to the ground-truth derived from the corresponding homography.

The remainder of this paper is organized as follows: Sect. 2 provides an overview of the keypoint detection and matching

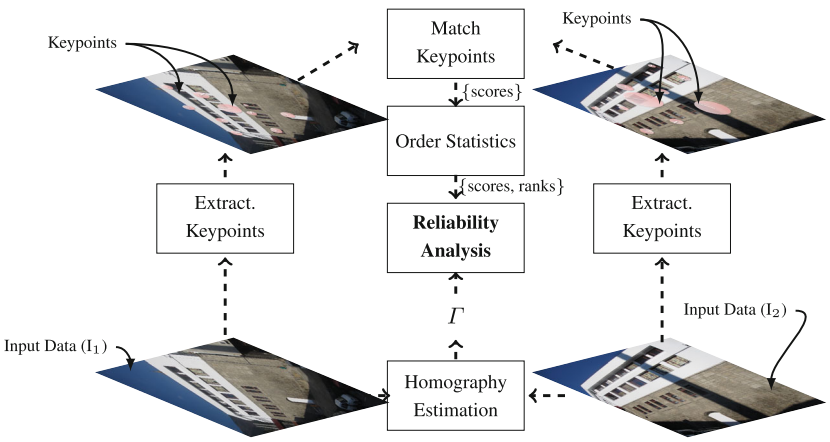

Fig. 1 Cohesive overview of the proposed evaluation framework. For each pair of images of the same scene (I_1 and I_2), a corresponding homography (Γ) was found. Keypoint detection and matching techniques were evaluated, based on the ground-truth provided by the corresponding homography

techniques considered. Section 3 describes our experiments and discusses the results. Finally, the overall conclusions are given in Sect. 4.

2 Evaluated techniques

Here we provide an overview of the techniques considered in this paper: Two of these describe algorithms for the complete detection/description/matching processing chain (SIFT and SURF), and the remaining (BRIEF and DAISY) focus exclusively on the description/matching phases. For the latter techniques, we considered different possibilities for the keypoint detection phase (Harris corner detection, visual saliency, SIFT and SURF detection). Table 1 summarizes the combinations of techniques evaluated.

2.1 Keypoint detection/description: SIFT

Proposed by Lowe [13], SIFT starts by obtaining a scale-space pyramid, approximated by the difference between Gaussian blurred images at different sigma values. Then, each element of the pyramid is considered a potential keypoint if it is a local extremum in a 3D neighborhood. For the purpose of accuracy, a 3D quadratic function is fitted to sample points $I(\mathbf{x})$ according to the Taylor expansion $I(\mathbf{x}) = I + \frac{\partial I}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\frac{\partial^2 I}{\partial \mathbf{x}^2}\mathbf{x}$. The exact location of the candidate keypoint is given by:

$$\hat{\mathbf{x}} = -\frac{\partial^2 I^{-1}}{\partial \mathbf{x}^2}\frac{\partial I}{\partial \mathbf{x}}. \tag{1}$$

Next, unstable extrema ($|I(\hat{\mathbf{x}})| < 0.03$) and edge responses $\left(\frac{\mathrm{tr}(H)^2}{\det(H)} < \frac{(r+1)^2}{r}\right)$ are rejected, where tr(.) and det(.) are the trace and the determinant, respectively, H is the Hessian matrix, and r is an empirically adjusted threshold. Finally, the magnitude of keypoints is given by the energy of the forward differences in the vertical and horizontal directions, and their orientation is given by the arctan of the proportion between these differences.

Regarding the description of keypoints, the concatenation of the magnitudes and orientation of gradients yields a feature vector. Next, keypoints are matched by the nearest neighbor strategy using the ℓ^2 norm. Having observed that many matches are due to background clutter, the authors compared the minimum distance between feature descriptors to the second smaller value. If this ratio is below a threshold, the correspondence is considered genuine.

Table 1 Combination of keypoint detection/description/matching techniques evaluated in this paper

Keypoint detection	Keypoint description	Keypoint matching
SIFT	SIFT	ℓ^2-norm
SURF	SURF	ℓ^2-norm
Harris, SIFT, SURF, visual saliency	BRIEF	ℓ^1-norm
Harris, SIFT, SURF, visual saliency	DAISY	ℓ^2-norm

2.2 Keypoint detection/description: SURF

In SURF [1], local keypoint detection starts by getting the Hessian matrix at different scales. Interest points are defined as local extrema in the 3D neighborhoods of the determinant of the Hessian matrix:

$$\det(H) \approx \frac{\partial^2 I}{\partial x^2}\frac{\partial I^2}{\partial y^2} - \left(\omega\frac{\partial I}{\partial xy}\right)^2, \tag{2}$$

being $\frac{\partial I}{\partial .}$ the image derivative with respect to a direction and ω a weight. To ensure invariance to rotation, authors maximized the energy of the Haar wavelet responses in horizontal and vertical directions at every point in a circular neighborhood.

In the description of each keypoint, a circular window with a radius proportional to scale is analyzed, and the responses to two orthogonal Haar wavelets are considered. For each region, four values are obtained: $\sum\frac{\partial I}{\partial x}$, $\sum\frac{\partial I}{\partial y}$, $\sum|\frac{\partial I}{\partial x}|$ and $\sum|\frac{\partial I}{\partial y}|$, and their concatenation yields the descriptor vector. In our experiments, matching between feature vectors was done according to the ℓ^2 norm. Similarly to the SIFT, a match was considered genuine only if the proportion between the top two nearest neighbors is below a threshold.

2.3 Keypoint description: BRIEF

Calonder et al. [2] considered that binary strings are efficient feature point descriptors. For each image pixel, a square region Ω of size s was analyzed and a binary test defined: $\tau(\Sigma, \mathbf{x}_1, \mathbf{x}_2) = 1$ if $p(\mathbf{x}_1) < p(\mathbf{x}_2)$; otherwise, $\tau(\Omega, \mathbf{x}_1, \mathbf{x}_2) = 0$, where $\mathbf{x}_.$ denotes an image location in Ω and $p(\mathbf{x})$ denotes the probability (pdf) of observing intensity values equal to $I(\mathbf{x})$ inside Ω. By choosing a set of $(\mathbf{x}_1, \mathbf{x}_2)$ locations (size t), the descriptor of the patch is given by the bit-string:

$$f(p) = \sum_{i=1}^{t} 2^{i-1}\tau(\Omega, \mathbf{x}_1, \mathbf{x}_2). \tag{3}$$

Different strategies were considered to obtain the spatial arrangements of the binary tests in Ω, concluding that performance is optimized when (X, Y) are independent and identically distributed random variables that follow a Gaussian distribution centered at 0 with variance of $s^2/25$. Finally, descriptors were matched according to the ℓ^1 norm.

2.4 Keypoint description: DAISY

Tola et al. [21] introduced a local image descriptor robust to photometric and geometric transformations. For every candidate keypoint, the authors drew circular regions of interest centered at each point with increasing radii, from which a number of orientation maps were extracted. The authors obtained these maps G for patches Ω of different sizes:

$$G_o^{\Omega} = G_{\Omega} * \left(\frac{\partial I}{\partial o}\right)^{+}, \tag{4}$$

where I is the input image, o is the orientation, and $(a)^{+} = \max(a, 0)$. Depending on the number of layers and of the radii of circular paths, the number of orientations and of the accumulated gradients is stored in histograms. The DAISY descriptor results from the concatenation of these histograms.

2.5 Keypoint detection: Harris

The Harris corner detector [7] is a well-known technique for locating interest points. It is based on the local autocorrelation function, measuring how similar a signal is to the same signal shifted toward different directions:

$$c(x, y, \Delta x, \Delta y) = \sum_{(u,v)\in\Omega} \Big(I(u, v) - I(u + \Delta x, v + \Delta y)\Big)^2. \tag{5}$$

The shifted version I is obtained by the first-order expansion of the Taylor series $I(u + \Delta x, v + \Delta y) \approx I(u, v) + [\frac{\partial I}{\partial x}(u, v) + \frac{\partial I}{\partial y}(u, v)] \cdot [\Delta x, \Delta y]^{\mathrm{T}}$. $c()$ is estimated by a quadratic function $[\Delta x, \Delta y]\, Q(x, y)\, [\Delta x, \Delta y]^{\mathrm{T}}$, given by:

$$Q(x, y) = \begin{bmatrix} \sum_{\Omega} \frac{\partial^2 I}{\partial x^2}(u, v) & \sum_{\Omega} \frac{\partial I}{\partial xy}(u, v) \\ \sum_{\Omega} \frac{\partial I}{\partial xy} & \sum_{\Omega} \frac{\partial^2 I}{\partial y^2}(u, v) \end{bmatrix} \tag{6}$$

Interest points correspond to positions with high values for both eigenvalues of Q.

2.6 Keypoint detection: visual saliency

Itti et al. [10] proposed a biologically based algorithm based on the feature integration theory. A dyadic Gaussian pyramid was built, and features extracted by a center-surround operation $I(c, s) = |I(c) \ominus I(s)|$, being $I(c)$ and $I(s)$ representations of the data at coarse and fine scales. Authors obtained

Fig. 2 Examples of the images used in our experiments. They regard a broad spectrum of scenes (e.g., buildings, objects, forests and paintings) and were acquired in non-controlled conditions with joint variations in terms of scale, rotation, perspective, occlusion and lighting conditions

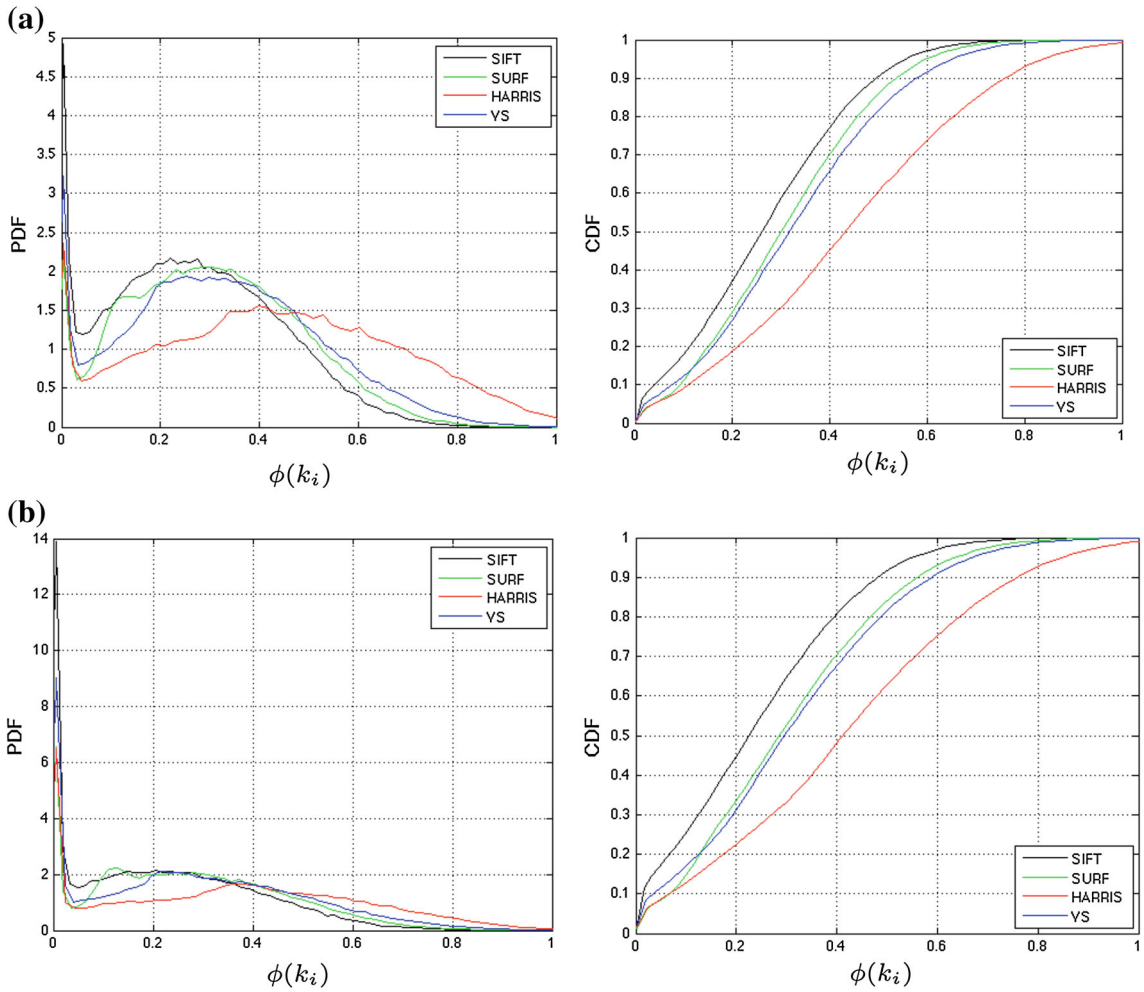

Fig. 3 Pdf and cdf values observed for the BRIEF (**a**) and DAISY (**b**) descriptors, with respect to the four keypoint detection methods considered

different linear combinations of these center-surround operators using red-, green- and blue-intensity data channels. The saliency map is the arithmetic mean of the three maps $\frac{1}{3}\big(N(I) + N(C) + N(O)\big)$, where $N(I)$, $N(O)$ and $N(C)$ denote the normalized intensity, orientation and color maps, respectively.

3 Experiments and discussion

3.1 Data sets and ground-truth

To ensure the reproducibility of the results, the data set and the annotation files used in this evaluation are publicly available.[1] 686 images were used, acquired using a *Canon Digital Ixus 99 IS* camera and covering 65 scenes. Images were resized from 4,000 × 3,000 to 800 × 600 pixels using bilinear interpolation, converted into grayscale and stored in bitmap format. All scenes contain non-deformable objects, captured from varying distances and 3D angles under non-controlled lighting conditions (Fig. 2).

For every pair of images of the same scene, a set of corresponding points were manually annotated. These points were used to obtain a homography Γ that maps each position in one image into the other:

$$[x_1, y_1, z_1]^{\mathrm{T}} = \Gamma[x_2, y_2, z_2]^{\mathrm{T}} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}, \tag{7}$$

where (x_1, y_1, z_1) and (x_2, y_2, z_2) denote the homogenous coordinates of points in both images. Γ is non-singular and homogeneous, with eight degrees of freedom. Two linear equations were derived for each point correspondence k_i:

$$\begin{cases} x_1(h_{31}x_2 + h_{32}y_2 + h_{33}) = h_{11}x_2 + h_{12}y_2 + h_{13} \\ y_1(h_{31}x_2 + h_{32}y_2 + h_{33}) = h_{21}x_2 + h_{22}y_2 + h_{23} \end{cases}. \tag{8}$$

[1] http://www.di.ubi.pt/~hugomcp/doc/evaluationDescriptors.zip.

After algebraic manipulation, (8) can be rearranged as follows:

$$\begin{bmatrix} x_2 & y_2 & 1 & 0 & 0 & 0 & -x_1x_2 & -x_1y_2 & -x_1 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -y_1x_2 & -y_1y_2 & -y_1 \end{bmatrix} \Gamma = 0. \quad (9)$$

Having at least four correspondences, the matrix Γ was derived by solving the corresponding system of linear equations. Hartley and Zisserman [8] provide additional details on this technique.

3.2 Performance evaluation

Let (I_1, I_2) be a pair of images of the same scene. When running a keypoint detection/matching algorithm, a set of correspondences between points in I_1 and I_2 are reported. Let k_i denote the ith correspondence and d_i denote the corresponding distance between the feature descriptors, i.e., $d_i = \xi\big(f(x_1, y_1), f(x_2, y_2)\big)$, being (x, y) the positions on both images, $f : \mathbb{N}^2 \rightarrow \mathbb{R}^n$ the feature descriptor and ξ a distance function. Let Γ be a homography that maps every point of I_2 into I_1. For each k_i, the error function $\phi(k_i) : \mathbb{N} \rightarrow [0, 1]$ has an inverse correspondence to the goodness of k_i:

$$\phi(k_i) = \frac{||(x_1, y_1) - \Gamma(x_2, y_2)||_2}{\sqrt{W^2 + H^2}}, \quad (10)$$

where $\Gamma(x_2, y_2)$ denotes the location of the keypoint (x_2, y_2) on the first image, when transformed by Γ. (W, H) are the image width and height and act as normalization terms.

In all our subsequent analysis, the term *performance* refers to the reliability of keypoint correspondences reported by matching algorithms, i.e., *high performance* corresponds to cases where most correspondences are genuine, and is equivalent to $\phi(k_i) \approx 0, \forall k_i$. Oppositely, *poor performance* is equivalent to cases where values of $\phi(k_i)$ are large, i.e., there is a poor agreement between the keypoint correspondences reported by the algorithm and the ground-truth data.

3.3 Performance of keypoint correspondences

As DAISY and BRIEF do not comprise the keypoint detection phase, their performance was assessed with four detection techniques: the Harris, visual saliency, SIFT and SURF detectors. Figure 3 gives the pdf and the cumulative (cdf) distribution functions for values of $\phi(k_i)$. The SIFT keypoint detection returned the best results for both the BRIEF and DAISY descriptors (mean values of $\phi(k_i) = 0.2872$ for BRIEF and $\phi(k_i) = 0.2348$ for DAISY). SURF and visual saliency provided similar results, and the poorest performance was observed for the Harris corner detector. Henceforth, all the experiments about BRIEF and DAISY include the SIFT keypoint detection phase.

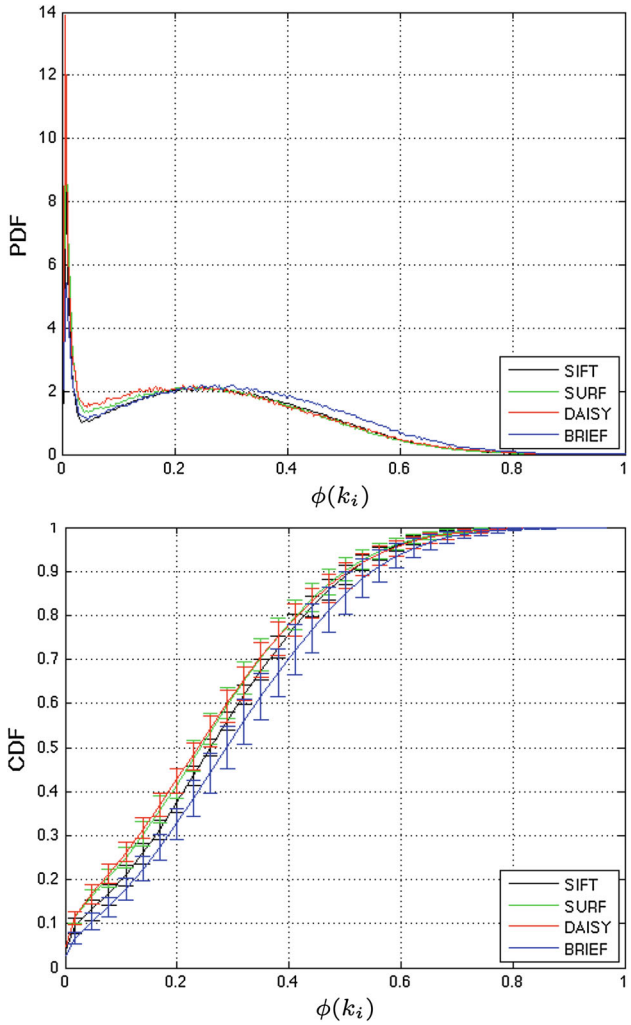

Fig. 4 Comparison between the performance of BRIEF, DAISY, SIFT and SURF techniques. The horizontal lines denote the best and worst performance observed at each operating point

Table 2 Mean matching errors observed for the BRIEF, DAISY, SIFT and SURF techniques

Keypoint detection/matching	Mean error $\phi(k_i)$
BRIEF	0.2872
DAISY	0.2348
SIFT	0.2685
SURF	0.2400

Figure 4 provides the distributions of the $\phi(k_i)$ values for the four algorithms evaluated. The upper plot gives the pdf, and the bottom plot gives the cdf values. Error bars represent 95 % confidence intervals calculated via bootstrapping. On the basis of these results, the major conclusions were as follows: (1) Globally, algorithms got close performance values, and the confidence intervals intercept in most regions of the space; (2) DAISY and SURF got the slight best results, which is particularly evident in the cdf plot; (3) SIFT got the

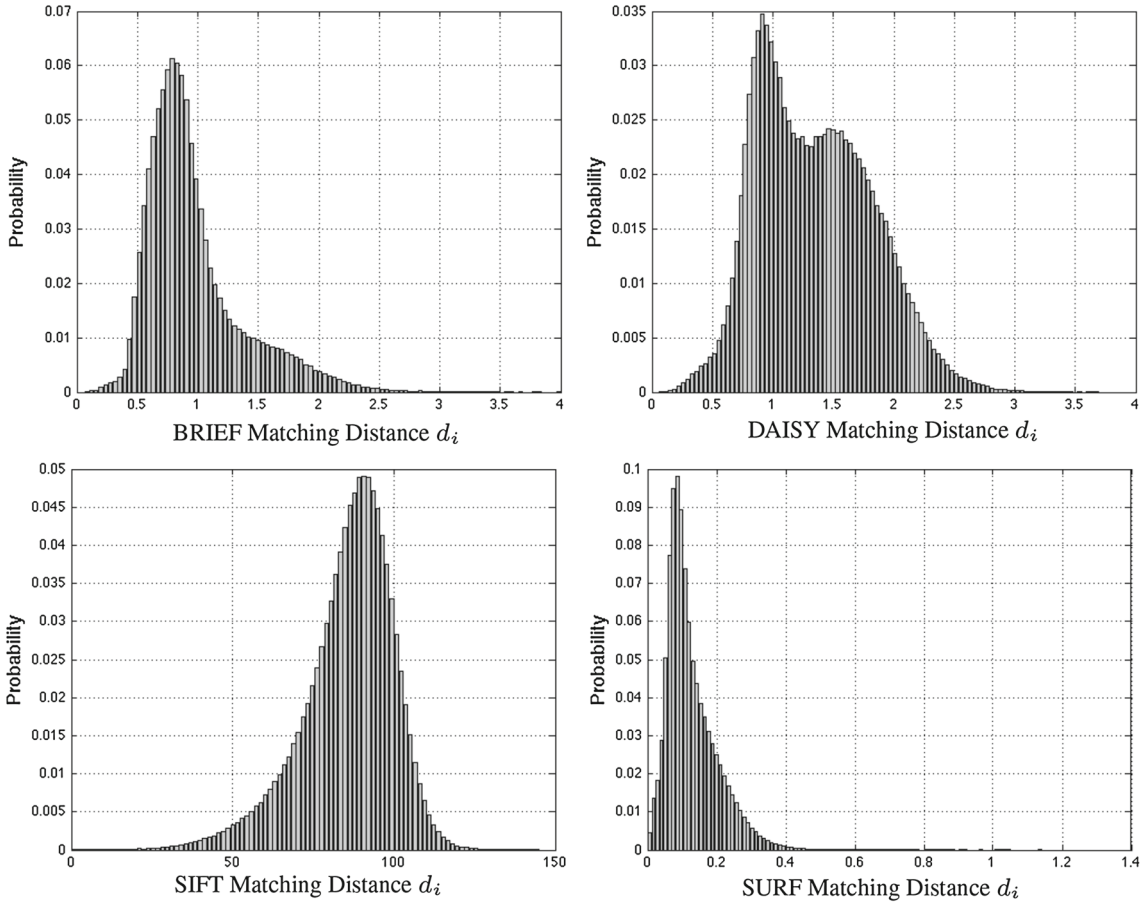

Fig. 5 Histograms of the matching distances d obtained for the SIFT, SURF, DAISY and BRIEF algorithms

third rank in terms of performance; and (4) BRIEF exhibited the poorest performance. It is noteworthy that all methods produced bimodal distributions, with the highest density around $\phi(k_i) = 0$ (i.e., when the keypoints were successfully matched). Table 2 summarizes the performance values in terms of the mean error.

Having concluded that none of the techniques consistently outranked the others and can be simply considered *the best*, our subsequent analysis was focused on the reliability of each technique with respect to the feature distances between the matched keypoints. A reasonable hypothesis is that the reliability of the correspondences reported varies with respect to the values of d_i, i.e., the *matching distance*. Intuitively, very low distances augment the confidence on the correspondence, whereas high distances have high pdf of regarding a spurious keypoint correspondence.

Figure 5 gives the pdfs of the matching distances d_i. Having DAISY as exception, all the techniques yielded unimodal distributions. SIFT was approximated by a negative skewed Gaussian curve, whereas BRIEF and SURF have a strong positive skew and were modeled by log-logistic distributions. Finally, the distribution of DAISY scores was modeled by a Gaussian mixture model.

3.4 Relationship between matching distances, order statistics and performance

To assess the relationship between the d_i values, their order statistic and the performance $\phi(k_i)$, two statistical tests were conducted, with *null* hypothesis:

- $\mathcal{H}_0^d$: There is no relationship between the matching distances d_i and performance $\phi(k_i)$.
- $\mathcal{H}_0^o$: There is no relationship between the order statistic of matching distances $o(d_i)$ and performance $\phi(k_i)$.

Table 3 Statistical tests of the linear correlation between matching distances d, their order statistics and performance $\phi(k_i)$

Method	Test	ρ	t	Status
BRIEF	$\mathcal{H}_0^d$	0.0541	69.27	✘
BRIEF	$\mathcal{H}_0^o$	0.1473	190.41	✘
DAISY	$\mathcal{H}_0^d$	0.0784	141.42	✘
DAISY	$\mathcal{H}_0^o$	0.1546	281.40	✘
SIFT	$\mathcal{H}_0^d$	0.1797	838.94	✘
SIFT	$\mathcal{H}_0^o$	0.0858	404.80	✘
SURF	$\mathcal{H}_0^d$	0.0799	219.64	✘
SURF	$\mathcal{H}_0^o$	0.2346	661.30	✘

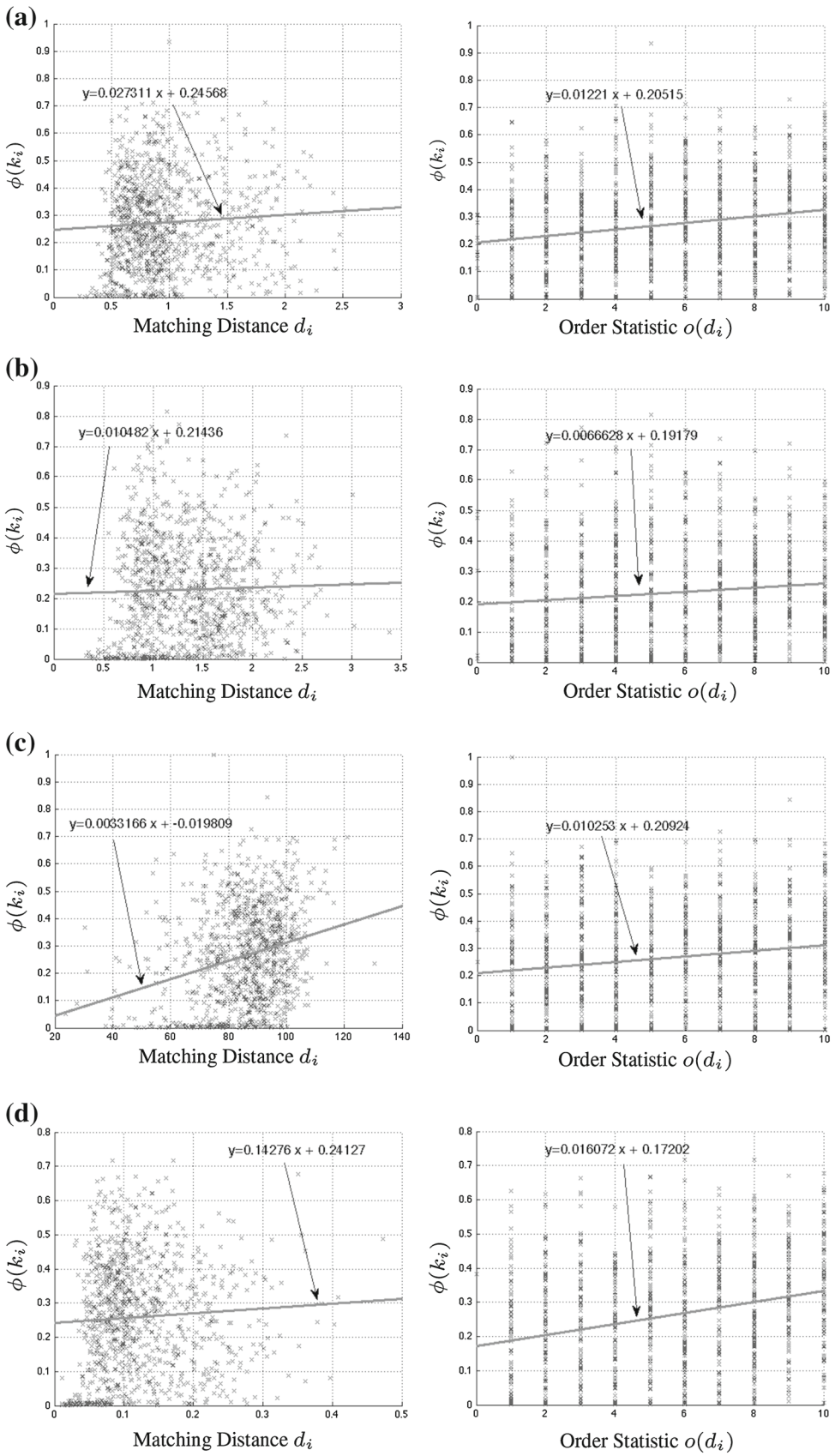

Fig. 6 Scatter plots of the linear relationship between matching distances d, their ranks and performance of SIFT (**c**), SURF (**d**), DAISY (**b**) and BRIEF (**a**) strategies

Pearson's correlation coefficient is a measure of linear correlation and is given by:

$$\rho(S_1, S_2) = \frac{n\sum s_1 s_2 - \left(\sum s_1\right)\left(\sum s_2\right)}{\sqrt{\left(n\sum s_1^2 - \left(\sum s_1\right)^2\right)\left(n\sum s_2^2 - \left(\sum s_2\right)^2\right)}}, \tag{11}$$

where S_1 and S_2 are the two compared statistics of size n. $\rho(.,.)$ has a t-distribution with $n-2$ degrees of freedom, and the test statistic is given by $t = \rho\sqrt{\frac{(n-2)}{1-\rho^2}}$, at $\alpha = 0.01$ significance level. Accordingly, if $|t| > 2.81$, the *null* hypothesis is rejected.

As given in Table 3, the observed t values were much larger than the critical value and, for all cases, the *null* hypothesis was rejected. Hence, it can be concluded that *there is a solid relation between the performance of keypoint correspondences and both the distances between feature descriptors and their order statistic*.

Figure 6 quantifies the strength of the relationships between performance ($\phi(k_i)$ column in the vertical axis), matching distances d_i (horizontal axis in the left subfigures) and their order statistic $o(d_i)$ (horizontal axis in the right subfigures). The straight lines are the least-squares fitting: Higher slopes denote stronger relationships. The strongest impact on performance was observed for SIFT with respect to the matching distances, even though it is noteworthy that the relation between performance and ranks of matching distances was generally stronger than the relation between performance and matching distances.

The above conclusion is confirmed by the analysis of Fig. 7, showing these relations in a visual way (the independent variables appear quantized in normalized intervals $\{1, \ldots, 10\}$). In these plots a linear and monotonous variation in the dependent variable ($\phi(k_i)$) with respect to the independent variable is more evident in the bottom plot (order statistic) than in the upper plot (matching distances). In this case, matching distances appear to be related to performance in a highly nonlinear way.

This observation was regarded as one of the major findings reported in this paper and has a clear insight: Order statistics preserve contextual information about the set of matching distances and—as such—provide a more reliable confidence measure for keypoint correspondences than matching distances.

3.5 Which technique is the best?

In this section we address the question about the *best* method to use in a practical scenario, given some prior information about the expected matching distances d_i. For such, we analyzed the joint relationship between performance $\phi(k_i)$, matching distances and their order statistic. The joint pdf mass functions were obtained by $P\left(d_i, o(d_i)\right) = P\left(d_i|o(d_i)\right).P\left(o(d_i)\right)$, being d_i and $o(d_i)$ quantized in $\{1, \ldots, 10\}$ intervals.

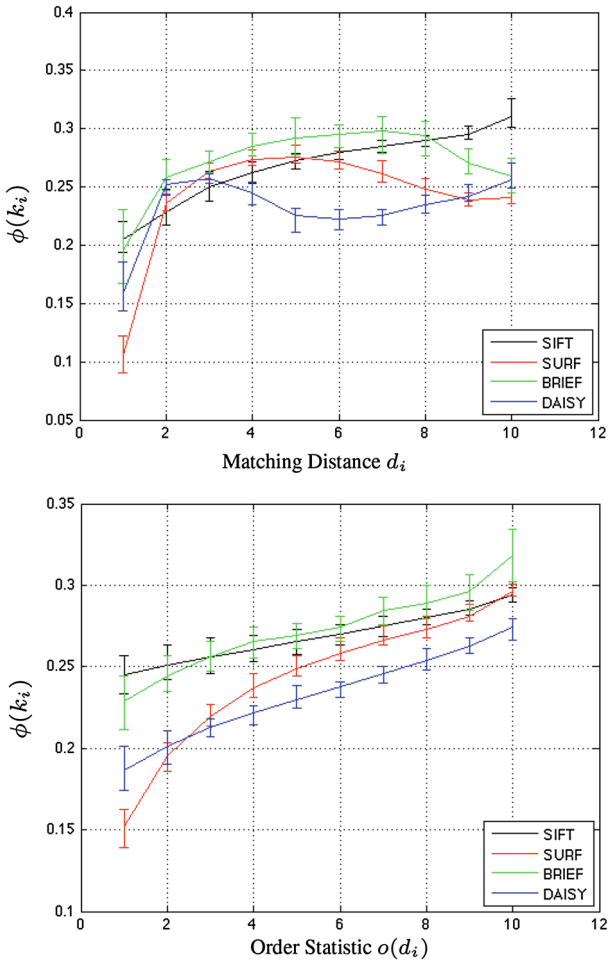

Fig. 7 Performance comparison for the SIFT, SURF, DAISY and BRIEF descriptors with respect to rank order (*bottom figure*) and matching distance (*upper figure*)

Figure 8 plots the joint relationships: Especially for the DAISY and SURF methods, the order statistics have a notably greater impact on performance than the matching distances. BRIEF exhibited the most homogeneous levels of performance and is the least susceptible to variations in performance with respect to matching distances and order statistics. Finally, an evident decline in the performance of SIFT was observed in one extreme of the performance range, enabling us to conclude that its reliability decreases significantly for high matching distances.

For comprehensibility, Fig. 9 summarizes the results and gives in different colors the *best method*, with respect to a range of matching distances and their order statistic. Note that each method outperforms all others at some operating range, turning evident that too simplistic conclusions (e.g., "*method A is better than method Y*") are erroneous. Instead, we concluded that:

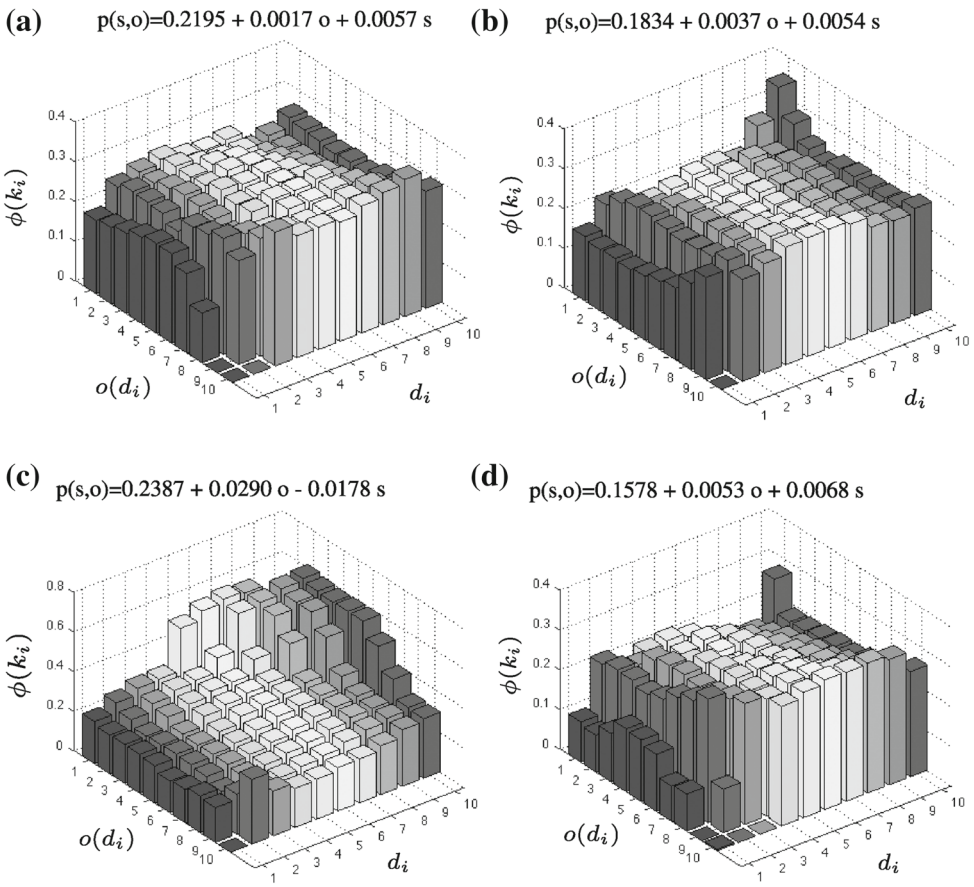

Fig. 8 Joint relationship between the matching distances d_i and their order statistic $o(d_i)$ for performance in the SIFT (**c**), SURF (**d**), DAISY (**d**) and BRIEF (**a**) methods. Next to each plot, the equation of the *least-squares* first-degree polynomial. Note that with the exception of SIFT, for all remaining methods, the order statistic is a better predictor of reliability than the matching distance

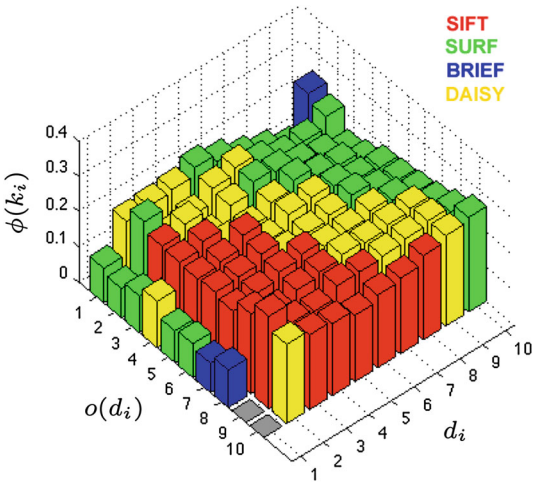

Fig. 9 Summary of the methods considered the *best*, with respect to a range of matching distances d_i and their order statistic $o(d_i)$. Colors denote the best observed performance at the corresponding operating range

– SURF is the most reliable technique if the matching distances are high and order statistics low. For practical cases, this corresponds to noisy acquisition environments, where a large heterogeneity between the matching distances is expected, also in cases where only the best keypoint correspondences are considered.
– In the other extreme of the performance space, SIFT is the most reliable technique if the matching distances are low and the order statistic is high (i.e., when the set of matching distances contains many low values). This indicates that SIFT is mostly suitable for good-quality environments, in which the pdf of obtaining low matching distances increases. Also, SIFT is the best technique in cases where a high proportion of the keypoints correspondences are used.
– DAISY is the best technique for intermediate cases, i.e., when order statistics and matching distances have approximately the same values in the quantized intervals.
– BRIEF hardly can be considered the best option for practical scenarios, even though it is the less sensitive to variations in performance with respect to matching distances and their order statistic.

4 Conclusions

In this paper we selected four of the most well-known methods for detecting/matching keypoints in grayscale images

(SIFT, SURF, BRIEF and DAISY) and carried out a systematic evaluation procedure with several singularities: (1) a large set of *real-world* data was used, representing a broad range of scenes with joint variations in translation, scale, rotation, perspective and lighting conditions, and (2) a completely automated evaluation procedure was devised, avoiding that subjectivity biases the obtained results.

According to our observations, the major conclusions are as follows: (1) There is a statistically significant relation between the matching distances reported by algorithms and the performance of each technique, and (2) specifically, the order statistic of the matching distances is a better predictor of performance than the matching distance itself.

When comparing the results attained by the techniques tested, we observed that general statements such as "*method A is better than method B*" are too simplistic and should be avoided. Instead, we concluded that SIFT, SURF and DAISY are *the best* in specific ranges of the performance space: SURF is better for noisy acquisition environments and for cases where only a fraction of the reported keypoint correspondences is used; SIFT is better for good-quality environments, where a large number of reliable correspondences are expected; and DAISY outperforms both techniques for environments of intermediate quality. Finally, BRIEF was not considered the most effective for any practical scenario, but is the less sensitive to variations in performance with respect to matching distances and their order statistic.

In this scope, further work should comprise the objective comparison between the performance of grayscale and color-based keypoint detectors, which have been gaining in popularity in the last few years. Also, it will be particularly important to perceive the correlation between the responses given by both types of descriptors, enabling us to perceive the actual role of color in this type of tasks.

Acknowledgments The financial support given by "FCT-Fundação para a Ciência e Tecnologia" and "FEDER" in the scope of the PTDC/EIA/103945/2008 research project "NECOVID: Negative Covert Biometric Recognition" is acknowledged.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: SURF: speeded Up robust features. Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
2. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: BRIEF: computing a local binary descriptor very fast. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1281–1298 (2012)
3. Carneiro, G., Jepson, A.: Phase-based local features. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 282–296 (2002)
4. Duy-Nguyen, T., Wei-Chao, C., Gelfand, N., Pulli, K.: SURFTrac: Efficient tracking and continuous object recognition using local feature descriptors. In: Proceedings of the IEEE Conference On Computer Vision and. Pattern Recognition (2009). doi:10.1109/CVPR.2009.5206831
5. Hajek, J., Sidak, Z.: Theory of rank tests, 1st edn. Academic Press, New York (2000)
6. Hanajik, M., Ravas, R., Smiesko, V.: Interest point detection for vision based mobile robot navigation. In: Proceedings of the IEEE 9th International Symposium on Applied Machine Intelligence and Informatics, pp. 207–211 (2011). doi:10.1109/SAMI.2011.5738876
7. Harris, C., Stephens, M.J.: A combined corner and edge detector. In Proceedings of the Alvey Vision Conference, pp. 147–152 (1988)
8. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2003) ISBN: 0-521-54051-8
9. Huang, C., Chen, C., Chung, P.: Contrast context histogram: an efficient discriminating local descriptor for object recognition and image matching. Pattern Recognit. **41**, 3071–3077 (2008)
10. Itti, L., Koch, C., Niebur, E.: Model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
11. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. Proc. Int. Conf. Comput. Vis. Pattern Recognit. **2**, 506–513 (2004)
12. Liu, C., Yuen, J., Torralba, A., Freeman, W.: SIFT flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 978–994 (2011)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
14. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 257–264 (2003)
15. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
16. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast keypoint recognition using random ferns. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 448–461 (2010)
17. Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T.: A thousand words in a scene. IEEE Trans. Pattern Anal. Mach. Intell. **29**(9), 1575–1589 (2007)
18. Randen, T., Husoy, J.: Filtering for texture classification: a comparative study. IEEE Trans. Pattern Anal. Mach. Intell. **21**(4), 291–310 (1999)
19. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. Int. J. Comput. Vis. **37**(2), 151–172 (2000)
20. Tao, Y., Skubic, M., Han, T., Xia, Y., Chi, X.: Evaluating color descriptors for object and scene recognition. Computer **2**(2), 17–20 (2010)
21. Tola, E., Lepetit, V., Fua, P.: DAISY: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 815–830 (2010)
22. van de Sande, E., Gevers, T., Snoek, C.: Evaluation of color descriptors for object and scene recognition. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
23. van de Sande, E., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1582–1596 (2010)