# Approximated Overlap Error for the Evaluation of Feature Descriptors on 3D Scenes

Fabio Bellavia[1], Cesare Valenti[2],
Carmen Alina Lupascu[2], and Domenico Tegolo[2]

[1] Università degli Studi di Firenze, Dipartimento di Sistemi Informatici,
Via di Santa Marta 3, 50139 FI, Italy
bellavia.fabio@gmail.com
[2] Università degli Studi di Palermo, Dipartimento di Matematica e Informatica,
Via Archirafi 34, 90123 PA, Italy
{cesare.valenti,carmen.lupascu,domenico.tegolo}@unipa.it

**Abstract.** This paper presents a new framework to evaluate feature descriptors on 3D datasets. The proposed method employs the approximated overlap error in order to conform with the reference planar evaluation case of the Oxford dataset based on the overlap error. The method takes into account not only the keypoint centre but also the feature shape and it does not require complex data setups, depth maps or an accurate camera calibration. Only a ground-truth fundamental matrix should be computed, so that the dataset can be freely extended by adding further images. The proposed approach is robust to false positives occurring in the evaluation process, which do not introduce any relevant changes in the results, so that the framework can be used unsupervised. Furthermore, the method has no loss in recall, which can be unsuitable for testing descriptors. The proposed evaluation compares on the SIFT and GLOH descriptors, used as references, and the recent state-of-the-art LIOP and MROGH descriptors, so that further insight on their behaviour in 3D scenes is provided as contribution too.

**Keywords:** keypoint descriptors, descriptor evaluation, epipolar geometry, SIFT, LIOP, MROGH.

## 1 Introduction

### 1.1 Related Works

Keypoint extracted from images have been adopted as primitive parts with good results in many computer vision tasks, such as recognition [10], tracking [9] and 3D reconstruction [16]. Detection and extraction of meaningful image regions, named keypoints or image features, is usually the first step of these methodologies. Numerical vectors that embody the image region properties are successively computed to compare the keypoints according to the particular computer vision task.

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 270–279, 2013.

Different feature detectors have been proposed in the last decades including, but not limiting to, corners and blobs, invariant to affine transformations or scale and rotation only. Since this paper deals with feature descriptor, the reader may refers to [13] for a general overview.

After the keypoint is located, a meaningful descriptor vector to embody the characteristic properties of the keypoint support region is computed. Different descriptors have been developed, mainly divided in two categories: distribution-based descriptors and banks of filters [12]. In general, while the former kind of descriptors gives better results, the latter category provides more compact descriptors [12]. Banks of filters include complex filters, colour moments, the local jet of the keypoint and the differential operators, please refer to [12] for more details.

Distribution-based descriptors, also named histogram-based descriptors, divide the image patch into different areas and compute specific histograms for some image properties of each area. The final descriptor is given by the ordered concatenation of these histograms. The rank and the census transforms [20], which consider binary tests of the intensity of the central pixel against its neighbourhood, is the precursors of the histogram based descriptors. The recent BRIEF [4] descriptor can be considered as an extension of this kind of approach, obtained by the concatenation of binary tests on the intensity values between couples of pixels of the patch. To be mentioned are the spin image descriptor, the shape context and the geometric blur, please refer to [12].

One of the most popular histogram based descriptor is the SIFT (Scale Invariant Feature Transform) descriptor [11], given by the 3D histogram of gradient orientations on a Cartesian grid. SIFT has been extended in various ways since its first introduction, for instance GLOH (Gradient Local Orientation Histogram) [12] combines a log-polar grid with PCA (Principal Component Analysis). Overlapping regions using multiple support regions combined by intensity order pooling are used by MROGH (Multi Support Region Order Based Gradient Histogram) [5]. Recently, LIOP (Local Intensity Order Pattern) [19] uses the relative order of neighbour pixels to define the histogram.

Feature descriptors and detectors have been evaluated on different frameworks [3, 6, 7, 12–15, 17]. In the case of planar images, the repeatability index and the matching score [13] established de facto standards for evaluate feature detectors, while precision/recall curves are used in the case of descriptors [12]. The Oxford dataset [13] is considered the standard benchmark for evaluation on planar images [12–14], but an extension to the 3D space is not immediate [7] or is limited to planar object in the 3D environment [8]. Other evaluation methodologies use laser-scanner images [17] or structure from motion algorithms [3] to generate the ground-truth data, while epipolar reprojection on more than two images have been also applied to constrain the matches [15], as wells as indirect evaluations by object recognition tasks [21]. These methodologies may require complex and error prone setups [7] or would not apply to any 3D scenario [3, 8] or may not have a direct relation with the overlap error [6, 21].

### 1.2 Contributions

We propose a new framework based on the approximated overlap error described in [1] to evaluate feature descriptors on 3D datasets. The approximated overlap error is reported for clarity in Section 2. This measure represents an overlap between a planar approximation of the surfaces inside the feature patches, so that the evaluation in the 3D case can be made very similar to that obtained through the overlap error in the case of planar scenes [12]. The approximated overlap provides a meaningful measure which considers the feature shape and not only the feature centre as proposed in [15]. Only the fundamental matrix is required, so that no complex schemas [3,7,15,17], threshold setups [3,15], depth maps [3,7] or an accurate camera calibration [7,15] are needed. This allows to easily extend the dataset to further images.

The proposed framework is reported in Sect. 3. A minimal user interaction can be required to inspect some matches by hand, since false positive can occur as for other approaches based on the epipolar geometry [6,15], due to wrong matches lying on correct epipolar lines. An experimental analysis of the expected errors in the case of an unsupervised application of the method is reported in Sect. 3.2, which shows that results are equivalent in practice to the supervised case. Finally, an effective comparison of recent descriptors based on the proposed framework is reported in Sect. 3.3. We compared the SIFT [11] and GLOH descriptors [12], used as references, with the recent LIOP [19] and MROGH [5]. According to recent evaluations [5,14] these last descriptors outperform other novel descriptors. Conclusions and final discussions are reported in Sect. 4.

## 2 The Approximated Overlap Error

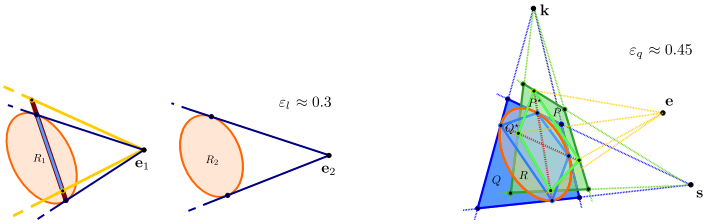
The working plan described by Mikolajczyk and Schmid [12] on the Oxford dataset can be considered a reference in the evaluation of descriptors in planar cases. For a stereo pair $(I_1, I_2)$ in the dataset, precision/recall curves on correct matches are extracted through the overlap error $\varepsilon$

$$\varepsilon(R_w, R_z) = 1 - \frac{R_w \bigcap \mathcal{T}_{2\to1}(R_z)}{R_w \bigcup \mathcal{T}_{2\to1}(R_z)} \tag{1}$$

where $R_w \in I_1$, $R_z \in I_2$ are the matched features and $\mathcal{T}_{2\to1}$ is the function that reprojects the feature $R_z$ from $I_2$ to $I_1$.

Precision/recall curves for a fixed overlap error $\varepsilon = 50\%$ are extracted for increasing distance values by using a match criterion chosen between the NN (Nearest Neighbour) and the NNR (Nearest Neighbour Ratio) matching [11]. The precision is defined as the fraction of correct matches according to the overlap error with respect to all matches observed so far [12] and the recall is the fraction of correct matches detected with respect to all the possible correct matches [12]. A good descriptor requires that the precision/recall curve increase rapidly and attains high recall values when it become stable [12].

The approximated overlap error $\varepsilon_q$ extends on planes the linear overlap error $\varepsilon_l$ introduced by Forseén and Lowe [6], which is briefly described for the sake of




(a) $\varepsilon_l$ (b) $\varepsilon_q$

**Fig. 1.** Overlap error approximations (best viewed in color)

clarity, see Fig. 1a. The tangency relation between epipoles and feature ellipses is preserved by perspective projection, since it is an incidence relation [18] (blue). By intersecting the line through the tangent points with the epipolar lines (yellow) of the corresponding tangent points on the respective ellipse in the other image, the linear overlap error $\varepsilon_l$ between the small (azure) and the wider (red) segments is obtained.

In [1] an extension to the linear overlap error measure $\varepsilon_l$ has been proposed, by observing that in order to compute the ground-truth fundamental matrix needed for $\varepsilon_l$, not only the correspondence between the epipoles is available, but also of fixed correspondences $(\mathbf{k}, \mathbf{s})$, $\mathbf{k}, \mathbf{s} \in C$, provided by additional hand taken points.

As shown in Fig. 1b, the tangent points of an ellipse patch $R$ (orange) define an inscribed quadrilateral $Q^\star$ (azure), while the tangent lines provide a circumscribed quadrilateral $Q$ (blue). The corresponding quadrilaterals $P^\star$ (light green) and $P$ (dark green) are obtained by projecting from the other images through the fundamental matrix as done for $\varepsilon_l$. With an abuse of notation, the area of the ellipse $R$ can be roughly approximated by the average area between $Q$ and $Q^\star$

$$R \approx \frac{Q + Q^\star}{2} \tag{2}$$

The final approximate overlap error $\varepsilon_q$ is defined as

$$\varepsilon_q = \frac{\varepsilon(Q, P) + \varepsilon(Q^\star, P^\star)}{2} \tag{3}$$

and it depends on the choice of the fixed correspondence pair $(\mathbf{k}, \mathbf{s})$. When epipolar and tangent lines are almost parallel the computation can suffer due to numerical instability. This issue can be resolved by introducing some heuristics [1].

The measure $\varepsilon_q$ is not symmetric so the maximal value obtained for the ordered stereo pairs $(I_1, I_2)$ and $(I_2, I_1)$ is assigned to the match. It represents an overlap between a planar approximation of the surfaces inside the feature patches. As other approaches based on the stereo epipolar geometry, it fails for some configurations. In particular, all feature ellipses which share the same cone

of epipolar lines with the correct matched feature are wrongly estimated as correct. This is however a stronger constraint with respect to consider only the epipolar distance from the keypoint centre. Furthermore, $\varepsilon_q$ gives better results than $\varepsilon_l$ when epipoles are inside the images [1] and can be directly related to the overlap error $\varepsilon$.

Since an exhaustive search should be done in order to find the best pair $(\mathbf{k}, \mathbf{s})$, the computational time $T(\varepsilon_q)$ depends quadratically on the cardinality $|C| = n$ of the fixed correspondence set $C$, i.e. $T(\varepsilon_q) = O(n^2)$. In order to limit $n$ for practical purposes, we introduce a new a greedy sampling strategy on the redundant correspondences of $C$ before the computation of $\varepsilon_q$ described in [1]. The fixed correspondences $c \in C$ used to compute the fundamental matrix are sorted in increasing order according to their epipolar errors. Both images are divided into a $m \times m$ grid and the hand taken correspondences $c$ are examined in turn. If any point of the correspondence $c$ falls in a grid cell of any of the images not covered by a previous retained correspondence, the correspondence $c$ is also retained. We experimentally verified that for a $24 \times 24$ grid size this strategy approximatively halves the number $n$ of the hand taken fixed correspondences, thus reducing the time required to about 30% without variations in the value of $\varepsilon_q$. About 20 minutes on a standard PC are required to get $\varepsilon_q$ on a stereo pair with roughly 800 matches and $n \approx 70$.

## 3 3D Evaluation

### 3.1 Setup

The evaluation on 3D images was done according to the freely available dataset[1] used in [1], see Fig. 3. The dataset consists of 10 different 3D scenes, with 3 images for each scenes so that a total of 30 stereo pairs are obtained. With respect to other evaluation methodologies, this comparison strategy does not require a complex setup [3,7,15,17], depth maps [3,7] or camera calibration [3,7,15]. Only the computation of a ground-truth fundamental matrix is required, so that the dataset can be easily extended by adding further images.

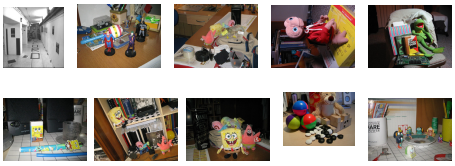


**Fig. 3.** Image dataset

**Table 1.** Evaluation strategy details

| | SIFT | GLOH | LIOP | MROGH | Average | All |
|---|---|---|---|---|---|---|
| FP (%) | 5.18 | 4.83 | 4.59 | 6.42 | 5.26 | 10.11 |
| Error (%) | 1.24 | 1.05 | 1.19 | 1.90 | 1.35 | 1.38 |
| Positives (%) | 24.02 | 21.67 | 26.02 | 29.52 | 25.30 | 13.66 |
| Checked (%) | 35.21 | 32.06 | 37.03 | 45.27 | 37.39 | 25.44 |
| Positives | 5712 | 5153 | 6188 | 7021 | 6018 | 10218 |
| Checked | 8375 | 7626 | 8807 | 10767 | 8893 | 19022 |
| Total | 23784 | 23784 | 23784 | 23784 | 23784 | 74784 |

The fundamental matrix is computed on more than 50 hand-taken correspondences for each stereo pair where the average epipolar error is of about 1 pixel.

[1] `http://www.math.unipa.it/fbellavia/dl/3d_eval.tar.bz2`

Note that the fundamental matrix computation does not require expert users, as small errors in the computation do not influence the shape area required by $\varepsilon_q$. This is due to the fact that not only the keypoint center is used, which would require an accurate estimation. We experienced that for higher epipolar reprojection errors in the test images, results on the approximated overlap error do not differ noticeably.

The HarrisZ detector [2], which selects robust and stable Harris corners in the affine scale-space, was used to extract keypoints. Previous evaluations [2] have shown that it is comparable with state-of-the-art detectors and provides better keypoints than Harris-affine. Furthermore, as noted in other work [5], although descriptors are influenced by detectors, the relative performances of the descriptors among different detectors are consistent [5, 7, 12, 15, 19]. The average number of keypoints extracted for the images of the dataset is about 800. For the evaluation of descriptors, SIFT and GLOH implementations by Mikolajczyk [12] are used, while the implementations available by the respective authors are used for LIOP [19] and MROGH [5].

The $\varepsilon_q$ measure can extract correct matches with a recall close to 100% and a precision between 70-98% when considering an overlap error threshold $t_{\varepsilon_q} < 1$ [1], i.e. matches are considered correct if they share only a minimal overlap region. As the threshold $t_{\varepsilon_q}$ decreases, the precision increases is similar way of the framework by Moreels and Perona [15].

The precision loss depends on the transformation applied to the scene as well as the uncertainly of the feature point on the epipolar line, as all the methodologies based on a ground-truth fundamental matrix. By using more than two images, as done in [15], this would be drastically reduced [1]. However, the recall of the correct matches is also reduced by about 30% [1,15] due to detector faults or occlusions, which can be unsuitable for testing descriptors. To deal with this issue, we compute $\varepsilon_q$ on the candidate matches and a first selection is done automatically by thresholding with $t_{\varepsilon_q} < 1$. The reduced subset of matches is then inspected by hands, by simply discarding matches not sharing at least a common point. This step requires a few minutes for each stereo pair by using the tool freely provided[3]. Furthermore, time decreases as more descriptors are analysed, as shared matches between descriptors do not need to be checked again.

The remaining matches with an approximated overlap error less than a predefined threshold, set to 0.5 as for the planar test [12], are only retained to obtain the final set of correct matches used as ground-truth in the evaluation. As discussed in the next section, to threshold directly to 0.5 without user interaction gives an average expected error rate close to 5%, similar to that found in [15] and it does not change the overall results of the evaluation.

Note that the approximated overlap error threshold was fixed at 0.5 just to make the experiment consistent with those on the planar test [12], but can be further reduced to improve the point location accuracy as other approaches [3, 15]. As a positive side effect, the time required by the user to check matches will be reduced too, since the number of the estimated correct matches decreases accordingly.

### 3.2 Evaluation Strategy Accuracy

Table 1 shows the estimated accuracy of the unsupervised alternative, i.e. when matches are threshold by $t_{\varepsilon_q} < 0.5$ without user inspection, in the case of the proposed evaluation. Details for each stereo pair are reported in the additional material[2]. In particular, wrong matches are reported with respect to all positive matches retained for $t_{\varepsilon_q} < 0.5$ in terms of false positives (FP) and as error rate with respect to the total number of matches. These values are reported by considering all stereo pairs in the dataset for each descriptor separately, their average value and all unique matches gathered from all descriptors, since detectors share matches.

The false positives rate is about 5% on average for a single descriptor. Among the 30 stereo pairs in the dataset, the approximated overlap error gives less than 4% of false positives for 23 image pairs, i.e. for about 75% of the image pairs in the dataset, see the additional material. The same considerations hold for the error rate which is about 1.3%. The obtained false positive rate does not change the relative ranks between descriptors, so that results obtained by the unsupervised evaluation are still consistent with the ground-truth results, as reported in the additional material.

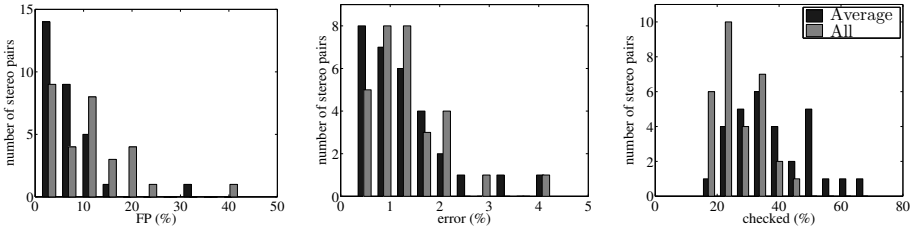
The false positive rate increases when considering all descriptors altogether. This accuracy decrease can be explained by considering that descriptors tend to share correct matches, which grow linear with the number of descriptors, while possible wrong matches increase quadratically. More wrong matches are present when the whole set of matches is considered, and the accuracy decreases accordingly.

The histograms of the false positives and of the error distribution reported in Fig. 4a-b show that the accuracy can considerably vary for different stereo pairs, in particular a considerable loss in the accuracy is present for the Kermit scene (rightmost image on the top row in Fig. 3), more details can be found in the additional material. This observation, should be taken into account in these kinds of evaluation.

Table 1 and Fig. 4c also report the number of matches that were checked for the supervised strategy in the proposed evaluation, i.e. all matches for which $t_{\varepsilon_q} < 1$. This value decreases from 37% for a single descriptor to 25% when we consider all descriptors since shared matches are taken into account. In the case of a direct threshold by $t_{\varepsilon_q} < 0.5$ to reduce the time spent, average values from 25% to 13% occur, corresponding to the positive matches in Table 1. Furthermore, matches to be checked are proportional to the correct matches, as the latter set is a subset of the former. For the same reasons, the matches to be checked proportionally decrease with the complexity of the stereo image pair.

Taking into account these observations, in order to add further new stereo pairs in the dataset, we suggest to check their accuracy for a reference descriptor and, if this value is reasonable, just follow the unsupervised evaluation for each descriptor to test.

[2] `http://www.math.unipa.it/fbellavia/dl/3d_eval_additional_material.pdf`




(a) (b) (c)

**Fig. 4.** Histograms of true positives (a), error rates (b) and checked matches (c) for the 30 stereo pairs in the dataset

### 3.3 Results

Fig. 5 shows the correct matches rate for each stereo pair and descriptor with respect to SIFT. We reported results for NNR matching with the $L_1$ distance, which is the best choice according to our experiments, see the additional material[4]. Histogram bars are clustered by scenes and then by stereo pairs. Inside a 3D scene, the stereo pair with the highest perspective distortion and occlusion achieves the lowest number of correct matches. The most challenging scene is represented by Kermit, where a high perspective distortion is present.

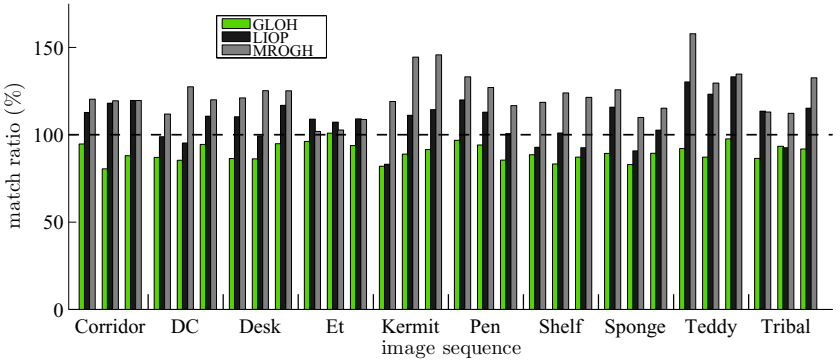



**Fig. 5.** Match ratio with respect to SIFT (best viewed in color)

GLOH performs slightly worse than SIFT with respect to other evaluations [12]. The difference in correct matches between SIFT, GLOH and rotation invariant descriptors is reduced from 50% in the planar case [5,19] to about 20% in 3D scenes. MROGH obtains in general the best results. Although LIOP does better than SIFT in most cases, it obtains a lower number of matches for some image pairs, because the local relative order of pixel intensities can suffer due to image discontinuities associated to perspective distortions in 3D image sequences.

Both the recently proposed rotational invariant LIOP and MROGH descriptors based on the intensity ordering pooling [5] are promising and can represent a new research direction toward the design of more efficient descriptors to avoid the bottleneck represented by the dominant orientation assignment [11] required by SIFT and GLOH.

## 4 Conclusions

This paper presents a new framework to evaluate feature descriptors on 3D datasets based on the recent approximated overlap error $\varepsilon_q$. This measure has a geometric mean which allows to uniform the evaluation on 3D scenes to the planar case.

The proposed framework, which does not require complex setups [3,6,7,12,13, 15,17], is freely available[1] and can be extended to include more images without relevant efforts. As other methods based on the epipolar stereo geometry [6,15], some false positive matches can occur, comparable to that obtained with other evaluation strategies, but no loss in recall is present since only stereo pairs are used, which makes this methodology appropriate to compare feature descriptors. The unsupervised use of the framework is still available, since the relative rank between the descriptors does not change.

By using the proposed evaluation strategy, a new comparison between the SIFT and GLOH descriptors with respect to novel rotational invariant approaches by intensity pooling, represented by LIOP and MROGH, is given. Descriptors based on the latter approach, especially MROGH, obtain better results, which show the validity on the rotational invariant method used.

Future works will include the extension of this evaluation scheme to further feature detectors and descriptors, with the addition of more images to the dataset.

## References

1. Bellavia, F., Tegolo, D.: New error measures to evaluate features on three-dimensional scenes. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part I. LNCS, vol. 6978, pp. 524–533. Springer, Heidelberg (2011)
2. Bellavia, F., Tegolo, D., Valenti, C.: Improving Harris corner selection strategy. IET Computer Vision 5(2) (2011)
3. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(1), 43–57 (2011)
4. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary robust independent elementary features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010)
5. Fan, B., Wu, F., Hu, Z.: Rotationally invariant descriptors using intensity order pooling. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(10), 2031–2045 (2012)

6. Forssén, P., Lowe, D.G.: Shape descriptors for maximally stable extremal regions. In: International Conference on Computer Vision. IEEE Computer Society Press (2007)
7. Fraundorfer, F., Bischof, H.: A novel performance evaluation method of local detectors on non-planar scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, p. 33. IEEE Computer Society Press (2005)
8. Gauglitz, S., Höllerer, T., Turk, M.: Evaluation of interest point detectors and feature descriptors for visual tracking. Int. J. Comput. Vision 94(3), 335–360 (2011)
9. Gil, A., Mozos, O.M., Ballesta, M., Reinoso, O.: A comparative evaluation of interest point detectors and local descriptors for visual SLAM. Machine Vision and Applications (MVA) 21(6), 905–920 (2010)
10. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169–2178. IEEE Computer Society Press (2006)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
13. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision 65(1-2), 43–72 (2005)
14. Miksik, O., Mikolajczyk, K.: Evaluation of local detectors and descriptors for fast feature matching. In: International Conference on Pattern Recognition (2012)
15. Moreels, P., Perona, P.: Evaluation of features detectors and descriptors based on 3d objects. International Journal of Computer Vision 73, 263–284 (2007)
16. Snavely, N., Seitz, S., Szeliski, R.: Modeling the world from internet photo collections. International Journal of Computer Vision 80(2), 189–210 (2008)
17. Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Computer Vision and Pattern Recognition (2008)
18. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer (2010)
19. Wang, Z., Fan, B., Wu, F.: Local intensity order pattern for feature description. In: IEEE International Conference on Computer Vision, pp. 603–610 (2011)
20. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 801, Springer, Heidelberg (1994)
21. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)