The effectiveness of detector combinations

4 authors, including:

Zhenghao Li
Chongqing University
9 PUBLICATIONS   **28** CITATIONS

SEE PROFILE

Andrew Y C Nee
National University of Singapore
510 PUBLICATIONS   **15,463** CITATIONS

SEE PROFILE

S K Ong
National University of Singapore
257 PUBLICATIONS   **7,733** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Manufacturing Sustainability in A Circular Economy View project

Special Session SS02 "Intelligent Manufacturing Systems" within the Conference "Manufacturing 2019" (May, 19-22, Poznan, Poland) View project

The effectiveness of detector combinations

Zhenghao Li[1, 2], Weiguo Gong[1*], A.Y.C. Nee[2], and S.K. Ong[2]

[1] *Key Lab of Optoelectronic Technology and System of Ministry of Education, Chongqing University*
Chongqing 400030, China
[2] *Department of Mechanical Engineering, Faculty of Engineering, National University of Singapore*
Singapore 117576, Singapore
*wggong@cqu.edu.cn

Abstract: In this paper, the performance improvement benefiting from the combination of local feature detectors for image matching and registration is evaluated. Possible combinations of five types of representative interest point detectors and region detectors are integrated into the testing framework. The performance is compared using the number of correspondences and the repeatability rate, as well as an original evaluation criterion named the Reconstruction Similarity (RS), which reflects not only the number of matches, but also the degree of matching error. It is observed that the combination of DoG extremum and MSCR outperforms any single detectors and other detector combinations in most cases. Furthermore, MDSS, a hybrid algorithm for accurate image matching, is proposed. Compared with standard SIFT and GLOH, its average RS rate exceeds more than 3.56%, and takes up even less computational time.

References and links

1. K. Mikolajczyk, and C. Schmid, "Scale & affine invariant interest point detectors," Int. J. Comput. Vis. **60**(1), 63–86 (2004).
2. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Vangool, "Comparison of affine region detectors," Int. J. Comput. Vis. **65**(1-2), 43–72 (2005).
3. K. Mikolajczyk, and C. Schmid, "Performance evaluation of local descriptors," IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005).
4. S. A. J. Winder, and M. Brown, "Learning local image descriptors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Minneapolis, United States, 2007), pp. 17–24.
5. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis. **60**(2), 91–110 (2004).
6. J. Matas, O. Chuma, M. Urbana, and T. Pajdla, "Wide-baseline stereo from maximally stable extremal regions," Image Vis. Comput. **22**(10), 761–767 (2004).
7. M. Donoser, and H. Bischof, "Efficient maximally stable extremal region (MSER) tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (New York, United States, 2006), pp. 553–560.
8. P. E. Forssén, "Maximally stable colour regions for recognition and matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Minneapolis, United States, 2007), pp. 1220–1227.
9. L. Trujillo, G. Olague, P. Legrand, and E. Lutton, "Regularity based descriptor computed from local image oscillations," Opt. Express **15**(10), 6140–6145 (2007).
10. T. Lindeberg, "Feature detection with automatic scale selection," Int. J. Comput. Vis. **30**(2), 79–116 (1998).
11. T. Tuytelaars, and L. Van Gool, "Matching widely separated views based on affine invariant regions," Int. J. Comput. Vis. **59**(1), 61–85 (2004).
12. T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," In *Proceedings of the European Conference on Computer Vision* (Prague, Czech Republic, 2004), pp. 345–457.
13. W. J. Schroeder, J. A. Zarge, and W. E. Lorensen, "Decimation of triangle meshes," Comput. Graph. **26**(2), 65–70 (1992).
14. J. S. Beis, and D. G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (San Juan, Puerto Rico, 1997), pp. 1000–1006.
15. Visual Geometry Group, "Affine covariant regions datasets," *http://www.robots.ox.ac.uk/~vgg/data/*.
16. A. Bowyer, "Computing Dirichlet tessellations," Comput. J. **24**(2), 162–166 (1981).

17. H. Øyvind, and D. Morten, "Triangulations and applications," in *Mathematics and Visualization*, G. Farin, H. C. Hege, D. Hoffman, C. R. Johnson, K. Polthier, and M. Rumpf, eds. (Springer, Berlin, 2006).
18. J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," IEEE Trans. Pattern Anal. Mach. Intell. **26**(1), 131–137 (2004).
19. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process. **13**(4), 600–612 (2004).
20. J. J. Corso, and G. D. Hager, "Coherent regions for concise and stable image description," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Diego, United States, 2005), pp. 184–190.
21. V. Lepetit, and P. Fua, "Keypoint recognition using randomized trees," IEEE Trans. Pattern Anal. Mach. Intell. **28**(9), 1465–1479 (2006).
22. V. Lepetit, and P. Fua, "Towards recognizing feature points using classification trees," in Technical Report IC/2004/74. (EPFL, 2004).
23. E. Rosten, and T. Drummond, "Fusing points and lines for high performance tracking," in *Proceedings of the IEEE International Conference on Computer Vision* (Beijing, China, 2005), pp. 1508–1515.
24. E. Rosten, and T. Drummond, "Machine learning for high-speed corner detection" in *Proceedings of the European Conference on Computer Vision* (Graz, Austria, 2006), pp. 430–443.
25. H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: speeded up robust features," in *Proceedings of the European Conference on Computer Vision* (Graz, Austria, 2006), pp. 404–417.
26. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," Comput. Vis. Image Underst. **110**(3), 346–359 (2008).
27. T. Liu, A. W. Moore, A. Gray, and K. Yang, "An investigation of practical approximate nearest neighbor algorithms," in *Advances in Neural Information Processing Systems*, L. K. Saul, Y. Weiss, and L. Bottou, eds. (MIT Press, Cambridge, 2005).
28. Z. Li, A. Y. C. Nee, S. K. Ong, and W. Gong, "Tampered image detection using image matching," in *Proceedings of the IEEE Computer Society Conference on Computer Graphics, Imaging and Visualization*, (Penang, Malaysia, 2008), pp. 174–179.

1. Introduction

Image matching, a method used to find the correspondence of two different images acquired from the same scene or objects at different intervals, or from different perspectives, is the key technology of many practical applications. It is currently one of the most active areas of research in computer vision. However, one can never achieve both high accuracy and low computational load at the same time. Hence, two mainstream research directions towards it have emerged. One is to maximize the accuracy on the premise of running in real-time, such as object tracking, visual navigation, and video retrieval; the other is to emphasize accuracy regardless of computation time and cost, e.g., image registration, 3D reconstruction, and panoramic scenes.

Adhering to the second direction, numerous scale and affine invariant detectors have been proposed. Meanwhile, many major evaluation methodologies have been conducted on these detectors, offering researchers good opportunities in making an exhaustive comparison. As a matter of fact, some light-weight detectors have already shown good performance in narrow baseline systems, but the number of reliable matches can drop sharply when there are large viewpoint and illumination changes, even when complex and high-end detectors are used. The authors explore a unique way utilizing detector combinations to unclog the performance bottleneck of a single detector, and obtain notable results. Compared with previous work [1–4], the authors have validated the effectiveness of detector combinations for the first time, and color space detectors have been integrated into their evaluation.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related work. In section 3, implementation details of the authors' evaluation are described. In section 4, experimental results are presented and discussed. A hybrid algorithm for accurate image matching is introduced in section 5. Lastly in section 6, conclusion and future work are presented.

2. Related work

A matching task can typically be divided into three modules. It begins with a feature detection module that detects features (interest points, edges, or regions) from the images, and is then followed by a description module where local features are converted to descriptor vectors that can be compared with the other descriptors in the final matching module. The first two

modules are both crucial to the accuracy, while in this paper, the authors mainly concentrate on the feature detection module.

Currently, the mainstream methods for feature detection include interest point detectors and region detectors. The authors do not intend to give a detailed description here, as similar work can be found in references [1–4].

In most literature, the performance is measured by the number of correspondences and the repeatability rate, i.e., the percentage of the points that are simultaneously presented in two images [1]. Though these two criteria indicate the potential matches, they cannot fully represent the overall performance. Schmid et al. and Winder et al. introduced the recall-precision [3] and Receiver Operating Characteristics (ROC) [4] for performance evaluation, and gave comprehensive reviews of the current descriptors. Their experimental results revealed that HSL (Hessian-Laplace) [1] and DoG (Difference of Gaussian) [5] outperform other interest point detectors, and MSER (Maximally Stable Extremal Region) [6,7] performs best among the region detectors on a wide range of test sequences. Note that the current algorithms and their evaluation criteria are mainly focused on gray-level images, but most test libraries are acquired through a color device. Therefore, the accuracy is definitely reduced due to the loss of color information.

3. Experimental details

Focusing on verifying the performance enhancement of detector combinations, evaluation is performed in the context of matching and recognition of images taken under different viewing conditions. The implementation details are given as follows.

3.1 Selection of candidate detectors

In accordance with the reviews mentioned in section 2 and the state-of-the-art literature [8,9], several representative detectors have been integrated into the testing framework. Five types are presented in this paper due to lack of space. The selected detectors are classified into two categories. One category is the interest point detectors, including DoG, HRL (Harris-Laplace), and HSL, and the others are the region detectors including MSER and MSCR (Maximally Stable Color Region) [8].

HRL detector uses the scale-adapted Harris function to locate its candidate interest points in scale space, while HSL detector locates its candidate interest points at the local maxima of Hessian determinant. Then, both of the two detectors perform scale selection using the scale-normalized Laplacian. DoG detector selects the local 3D extrema in a DoG pyramid as its interest points. DoG representation is a close approximation of LoG (Laplacian of Gaussian) [10], but DoG can significantly accelerate the computational process. MSER employs a watershed-like algorithm to derive the distinctive regions. MSCR is an extension of MSER in color space. This extension is achieved by looking at successive time-steps of an agglomerative clustering of pixels.

The classical LoG detector is not included because there is no observable difference compared to DoG when the image size is restricted to 800×600 pixels. Another two influential region detectors, EBR (Edge Based Region) [11] and IBR (Intensity extrema Based Region) [11], detect their regions based on the Harris corners or intensity extrema. Thus, there is little sense to test the combinations of them together with an interest point detector. The authors have also omitted the entropy-based salient regions [12] algorithm here, due to limited number of correspondences in the authors' evaluation (this is also reported by most researchers).

3.2 Experimental setup

All the possible detector combinations are compared, where each is combined with an interest point detector and a region detector. Therefore one can derive $3 \times 2 = 6$ types of candidate combinations from the selected detectors. Firstly, features are detected using a single detector. To obtain the location of a region detector includes three steps: 1) building an approximate ellipse of the region (not the actual region), 2) warping the elliptical region to the circular

region of constant radius to obtain scale and affine invariance, and 3) setting the center of the circle as the location of the region. The detected feature with the distance from itself to any other detected one within a threshold (a threshold of 3 pixels is chosen in the present experiments), is defined as a pair of redetected features. After removing one unreliable feature for every redetected pairs using vertex decimation [13], all the remaining ones are the features detected by the combination. The experimental results will be affected by the selection strategy of the unreliable features. Two strategies are tested. The first strategy keeps the features detected by the interest point detector and discards the features detected by the region detector. On the contrary, in the second strategy, only features detected by the region detector are kept. The data shown in this paper are obtained using the first strategy for the reason that point detectors tend to work more accurately than region detectors in locating a feature.

In the description module, the standard 128 dimensional SIFT descriptor [5], which is invariant to image scaling and rotation, is adopted. The SIFT descriptor is a 3D histogram of gradient locations and orientations. Each bin in this histogram is weighted by the gradient magnitude. In the matching module, the BBF (Best-Bin-First) approach [14] is used, which can reduce the search time through maintaining a priority queue of bins that are not visited.

There is a slight change to the existing evaluation scheme where the ratio of the nearest neighbor to the second-nearest one is also considered (the threshold is set to 0.6) for the sake of achieving more reliable and accurate correspondences. Finally, RANSAC, the effective way of eliminating spurious matches, should not be omitted.

3.3 Evaluation criteria

In order to generate comparable results, the number of correspondences and repeatability tests are covered into the present evaluation, and all the experiments are conducted using the standard test library [15] provided by the Visual Geometry Group (sample images are regulated to the size of 800×600). It is considered that the researchers' overriding concern is the general performance and the robustness to viewpoint and illumination changes. Thereby, the authors chose the 'Boat' sequence (scale change and rotation) to test the performance under usual view conditions, the 'Graffiti' sequence (multiplex transformation, manly viewpoint change) to test the stability in extreme cases, and the 'Leuven' sequence to testify the light invariability. In Fig. 1, the miniatures of first image in the three sequences are shown.

Fig. 1. Test library. Each sequence contains 6 images. From left to right: 1) 'Boat' sequence, 2) 'Graffiti' sequence, and 3) 'Leuven' campus sequence.

As discussed in section 2, though the number of correspondences as well as repeatability can reflect the performance of correct matching, neither of them can describe the degree of the matching error. Based on the authors' experience, some obviously wrong matches are hard to be eliminated by means of RANSAC or epipolar constraint in the wide baseline systems. However, the remaining ones would usually cause fatal errors in an application. Thus, the authors advocate that the number of correspondences and the extent of mismatch, rather than repeatability, should be emphasized, and bring forward an evaluation criterion called Reconstruction Similarity (RS).

In the RS test, the matched features are classified into two categories. The feature with the distance from itself to any other feature larger than a threshold (a threshold of 4 pixels is chosen) is defined as a non-overlapping feature. After performing an image matching task between two images, the redundant matches are firstly removed using vertex decimation, and

the Delaunay triangulation mesh is built [16,17] on the non-overlapping features. Next, a piece-wise affine deformation is performed on the mesh to reconstruct image 1 (Fig. 2) using the material from image 2 (Fig. 2). The last step is similarity measurement, in which, features are extracted using 2D-PCA [18], and the similarity is exported from a cosine classifier (hereinafter called PCACOS). Similarity is also assessed using a more rigorous index SSIM (Structural SIMilarity) [19] in section 5. The flow chart is shown in Fig. 2.
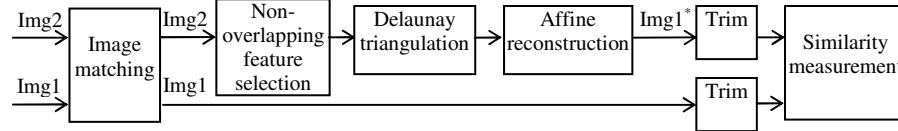
Fig. 2. Flow chart of RS test.

It is known that in the wide baseline systems, image deformation cannot be realistically approximated even when using full affine translation. However, according to the mechanism of RS test, if more accurate correspondences and a greater number of reliable non-overlapping matches can be obtained, then a higher RS rate can be achieved. Hence, the authors advocate using the RS criterion for its capability of fully representing the matching performance of a given algorithm.

4. Experimental results and discussion

In Fig. 3, the numbers of correspondences and the repeatability using single detectors-as well as the results for the same test using detector combinations are shown. The results of RS tests are depicted in Fig. 4. For clarity's sake, an average curve in red color with value is drawn in each chart.

It is known that the number of features is of particular importance in object recognition and tracking, where to correctly detect small objects in cluttered backgrounds requires at least three features [5]. It can be observed from Fig. 3 that the combination of two types of detectors can yield more reliable correspondences, especially in most challenging view conditions (large viewpoint and illumination change). The repeatability is not sacrificed or even increased except for the 'Leuven' sequence. In other words, such a combination has a desirable property that it ensures sufficient correspondences in small patches while maintaining high efficiency. Theoretically, HSL works more accurately than DoG due to the sensitivity of DoG to edges. However, the number of correspondences detected by HSL is usually less than DoG and HRL. As a matter of fact, inspired from [20], the authors have made an attempt to extend DoG to color space by performing the extrema detection on grey level, red-green channel, and blue-yellow channel (hereinafter called DoG3). Compared with the present combinations, more features can be detected using this extension, but the number of correspondence remains the same after removing the duplicated features.

Figure 4 gives a better comparison of the overall performance as the degree of error is taken into consideration in the RS test. In this group of charts, the combinations of detectors gain an overwhelming success compared to single detectors. This may also relate to the increased number of reliable matches. The combination of MSCR and DoG performs outstandingly compared to others, and it can be substituted by the combination of MSER and DoG in grey-level view condition due to equivalent performance.
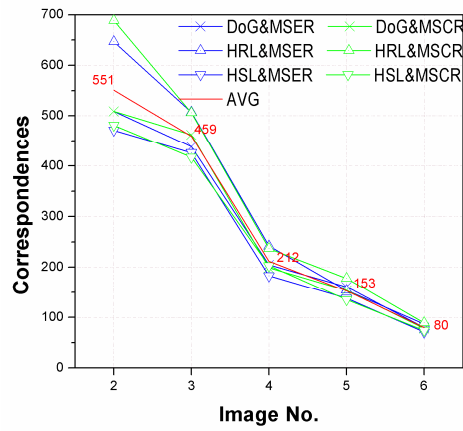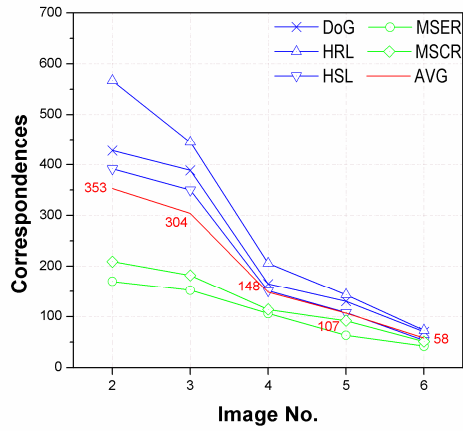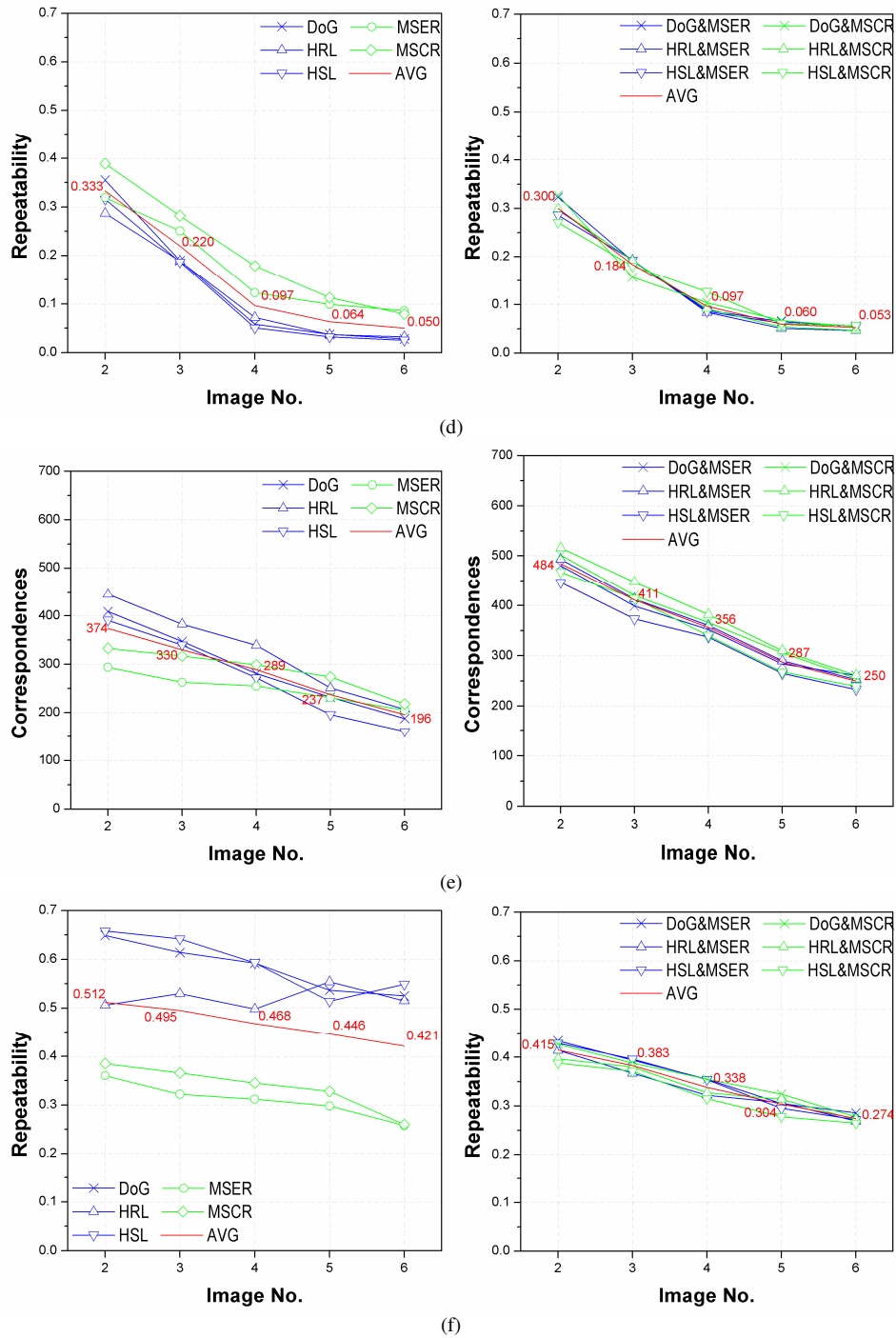
(a)
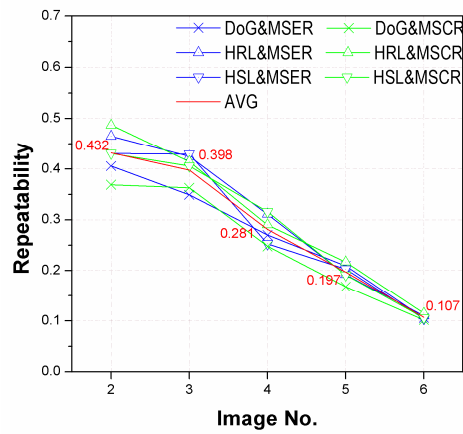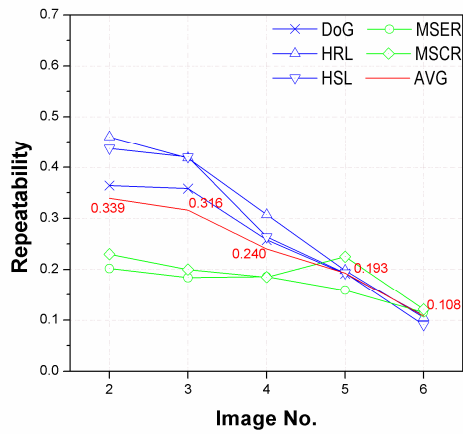
(b)
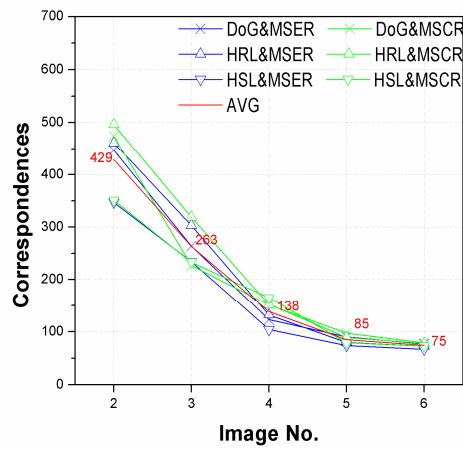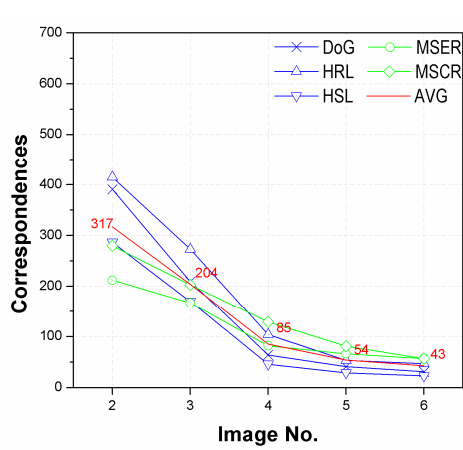
(c)

Fig. 3. Performance comparisons of different detectors. (a), (c) and (e) show the number of correspondences, and (b), (d) and (f) show the repeatability rate of 'Boat', 'Graffiti' and 'Leuven' campus sequence respectively.
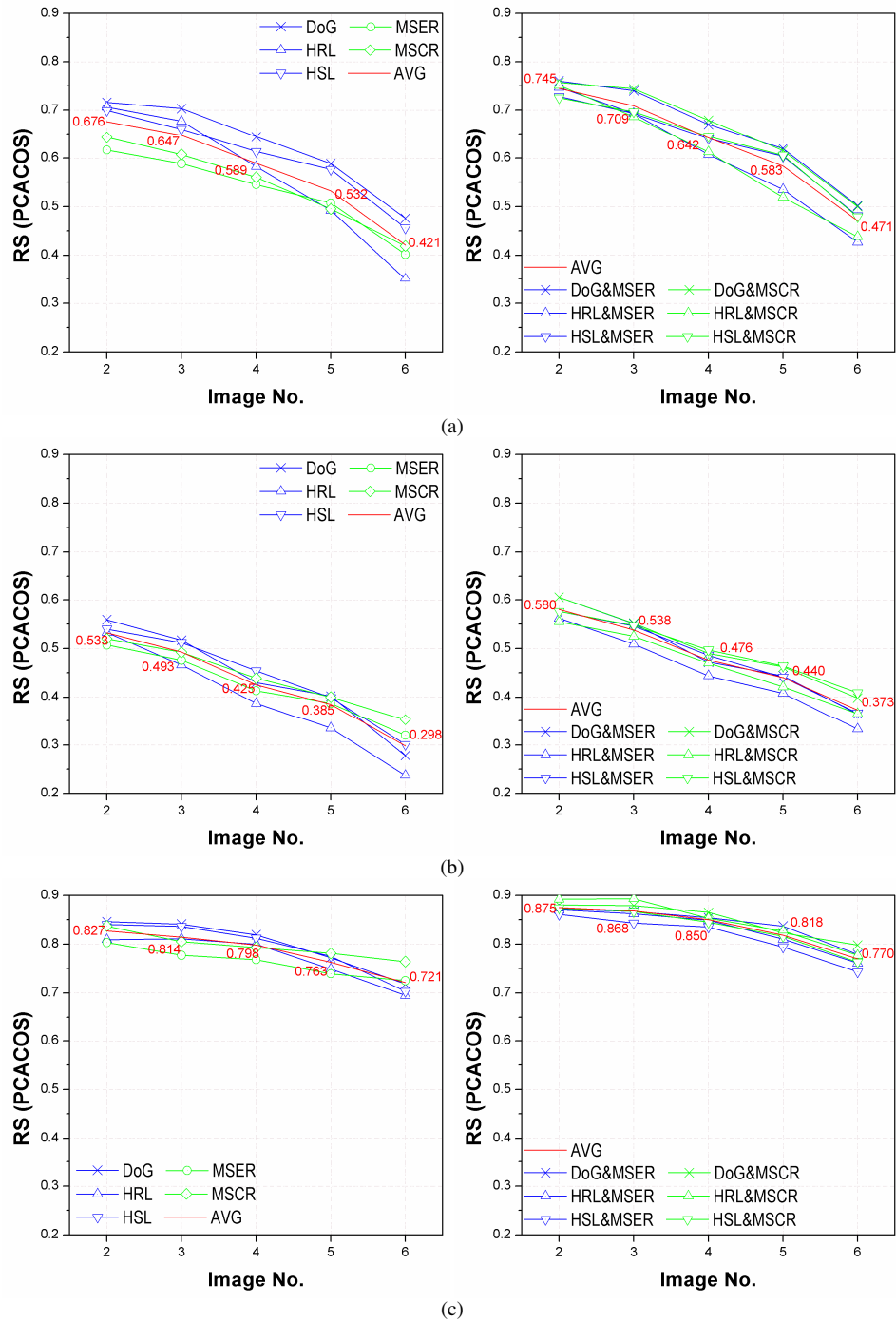
Fig. 4. Performance comparisons using RS test. (a), (b) and (c) show the RS rate of 'Boat', 'Graffiti' and 'Leuven' campus sequence respectively.

The following hints can be learnt from these charts.

(1) Though the largest number of correspondences is achieved by the combination of MSCR and HRL in most cases followed with the combination of MSER and HRL, these two combinations do not work well in the RS tests. This is because the Harris

function is inclined to be more sensitive to edge response and less distinctive than the Hessian determinant and DoG filter. Consequently, it usually causes too many erroneous matches, several of which are large-extent errors.

(2) Interest point detectors detect more correspondences than region detectors in narrow baseline vision, while region detectors show more stability where there are large viewpoint and light changes. Hence, the detector combination has indeed exerted the strength of individual detectors.

(3) Though HSL, etc., are recognized as discriminative approaches in feature detection, many "less distinctive" feature regions, which cannot be detected using interest point detectors, can still be well described by the SIFT descriptor, and can correctly find their correspondences, namely, selected detectors are not as powerful as SIFT descriptors. The intrinsic reason is that the additional procedures to eliminate the second-order effects, including skew, anisotropic scaling, are too rigid, resulting in a negative impact on the robustness, unless really large viewpoint changes are to be expected. Thereby, researchers may find the reason why modern detectors tend to locate keypoints by simply examining the intensities of certain pixels around the tentative locations [21–24], or adopting an approximate filter [25,26].

(4) Color information cannot be neglected for achieving high matching performance from the contrast between MSCR and MSER. The color version of the agglomerative clustering employed in MSCR can fully and effectively take advantage of the color information. However, the authors have not found a valid way to extend the interest point detector to function in color.

It should be noted that, as can be seen from the tests, MSCR outperforms MSER. Nevertheless, Forssén [8] reported that MSER does better in the 'Graffiti' sequence than MSCR. Though the ranking may vary slightly depending on the implementation details and the image content, the results obtained from the authors' evaluation are reliable.

5. MDSS matching algorithm

From the discussion in section 4, it can be concluded that the combination of MSCR and DoG could theoretically achieve the best performance, and it has been verified in the authors' evaluation. In order to generate a series of comparable results, the classical SIFT solution is adopted in feature detection and matching modules. However, with the development of searching algorithms, BBF is outdated.

We derived a hybrid method for practical applications. It is designated as MDSS for the reason that it is built up through combining the MSCR and DoG detectors, SIFT descriptor, and hybrid Spill Tree (SP-Tree) [27] search engine. The combination of MSCR and DoG, as well as the SIFT descriptor, are introduced in experimental setup section. However, in the feature matching module, BBF is replaced by the hybrid SP-Tree, a more efficient approximate nearest neighbor searching approach proposed by Liu et al. in 2005.

The hybrid SP-Tree is considered as a combination of the 'defeatist search' in the non-overlapping nodes and the SP-Tree in the overlapping nodes. The core idea of this approach is to properly set an overlapping buffer that the child-nodes can share their points, with which the time cost of back-tracking can be substantially reduced.

Figure 5 shows the matching results of image 1 and 4 in the 'Graffiti' sequence using SIFT and MDSS respectively. The performance comparison between MDSS and another five matching algorithms has been conducted, and the results are shown in Fig. 6 and Table 1. In this comparison, SIFT, GLOH and MSER are highly recognized matching solutions, and have become standards in benchmark tests. MDSS- is used to represent the gray level version of MDSS, in which MSCR is substituted for MSER. SIFT3 is an extension of SIFT in color space. The computational time listed in Table 1 is collected using an ordinary laptop with 2.1GHz Intel® Core™2 Duo CPU and 2GB RAM.

As it can be observed from Fig. 6, MDSS has always obtained the highest precision, closely followed by MDSS-. Table 1 indicates that the hybrid SP-Tree only takes up 37.32% of the computational time of BBF in the matching module. Finally, the time cost of MDSS matching is not longer than GLOH, but the precision maintains at 4.67% (PCACOS) and 3.56% (SSIM) higher than SIFT and GLOH on average.

(a)

(b)

Fig. 5. Matching results of image 1 and 4 in the 'Graffiti' sequence using SIFT (a) and MDSS (b).
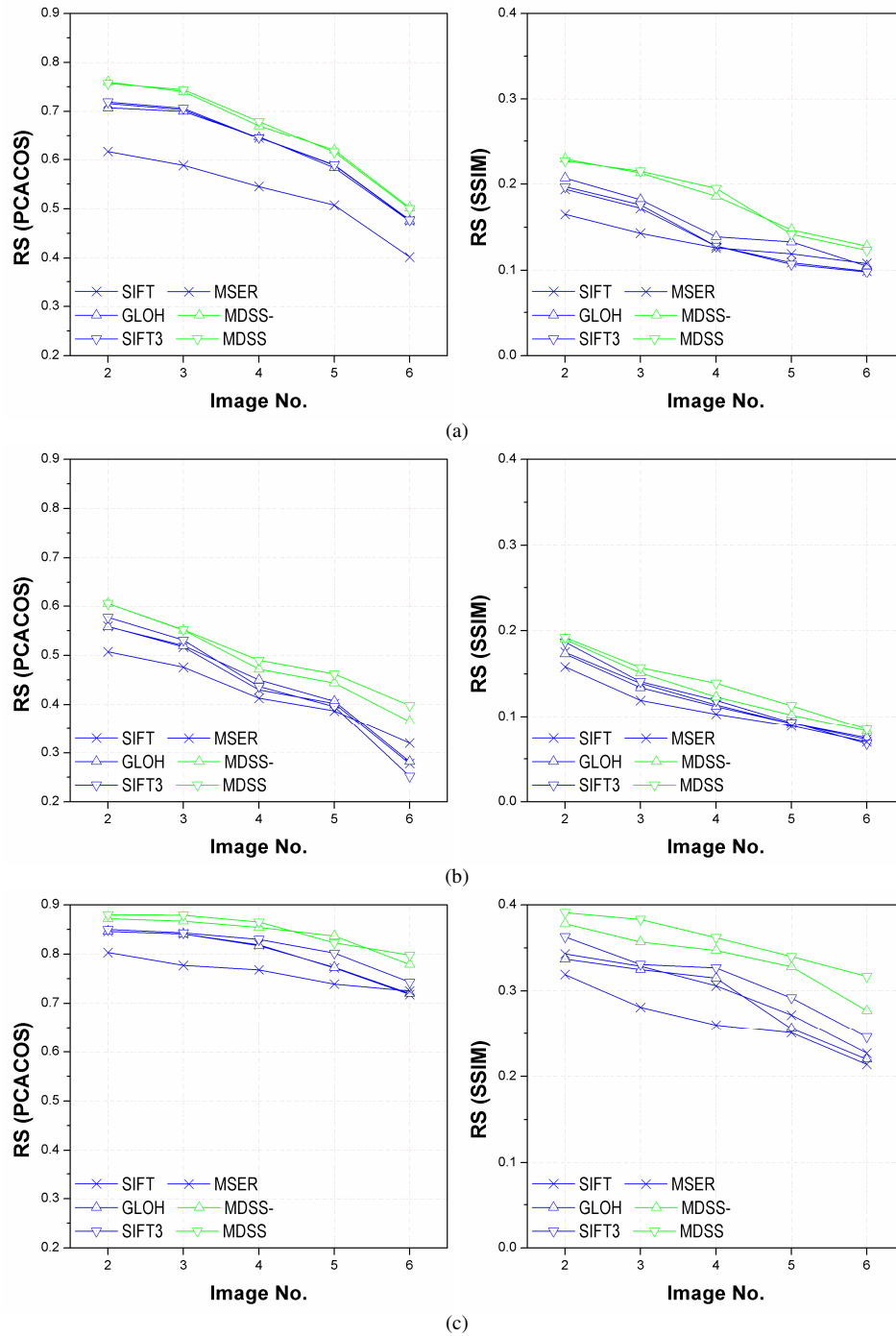
Fig. 6. Performance comparisons using RS test. **Rows**, from top to bottom: 1) 'Boat', 2) 'Graffiti', and 3) 'Leuven' campus sequence. **Columns**, from left to right: similarity computed by 1) PCACOS and 2) SSIM.

Table 1. Comparisons of different matching solutions

Algorithm	SIFT	GLOH	SIFT3	MSER	MDSS-	MDSS
Detector	DoG	DoG	DoG3	MSER	MSER + DoG	MSCR + DoG
Descriptor	SIFT	GLOH	SIFT	SIFT	SIFT	SIFT
Matching	BBF	BBF	BBF	BBF	hybrid SP-Tree	hybrid SP-Tree
Time in matching (sec/frame)	0.903	0.907	1.043	0.896	0.336	0.337
Total time (sec/frame)	2.312	2.507	4.308	1.575	2.389	2.411

6. Conclusions and future work

Though various unique detectors have emerged in recent years, few of them have made revolutionary breakthroughs. In this paper, the authors propose a feasible solution to enhance the overall matching performance. Obviously, this paper does not fully cover the existing algorithms, but the effectiveness of detector combination is revealed in the experimental comparisons. It was found that the combination of MSCR and DoG can substantially improve the performance of feature detection, and MDSS is proposed and presented for accurate image matching and object detection. The authors have applied the MDSS- to image tamper detection and achieved desirable results [28]. Moreover, the hints summarized in section 4 are worth taking close concern for other researchers. In addition, MDSS is not the fixed and ultimate version. If the strategy of locating features detected by a region detector is changed, for instance, choosing the two focus points of the approximate ellipse as the location instead of the center, more correspondences can be obtained.

In future, the authors will study a continuation, in which region detectors are used as a constraint to eliminate erroneous matches, in order to improve wide baseline performance. An effective way of utilizing color information in interest point detectors will also be further studied.