Scale Invariant Feature Transform with Irregular Orientation Histogram Binning

Yan Cui, Nils Hasler, Thorsten Thormählen, Hans-Peter Seidel

MPI Informatik, Saarbrücken, Germany

Abstract. The SIFT (Scale Invariant Feature Transform) descriptor is a widely used method for matching image features. However, perfect scale invariance can not be achieved in practice because of sampling artefacts, noise in the image data, and the fact that the computational effort limits the number of analyzed scale space images. In this paper we propose a modification of the descriptor's regular grid of orientation histogram bins to an irregular grid. The irregular grid approach reduces the negative effect of scale error and significantly increases the matching precision for image features. Results with a standard data set are presented that show that the irregular grid approach outperforms the original SIFT descriptor and other state-of-the-art extentions.

1 Introduction

The reliable matching of image features is a basic problem in computer vision applications, like 3D reconstruction from stereo images [1], structure-and-motion estimation [2], panorama generation [3], or object recognition [4]. Especially, if the change in 3D viewpoint between the images is large, the matching of the image features must be invariant to image transformations and illumination changes. Usually, the matching process can be divided into two steps. The first step is the detection of feature points (also called keypoints). In this step descriptive image regions are selected and their exact image position is determined. The second step is the keypoint correspondence analysis, where pairwise assignments of keypoints are determined based on local region descriptors (also called keypoint descriptors).

A well-established keypoint detector and descriptor is the Scale Invariant Feature Tranform (SIFT), which was published in 2004 by Lowe [5]. After detection and localization of keypoints in different scale space images, an orientation is assigned to each keypoint using local image gradients. Then a keypoint descriptor is assembled from the local gradient values around each keypoint using orientation histograms. In 2005, Mikolajczyk and Schmid [6] carried out a performance evaluation of local descriptors and concluded that the SIFT-based descriptor performs best.

In this paper, we suggest a modification, which differs from the original SIFT approach only in how the keypoint descriptor is assembled from the local gradient values around each keypoint. Instead of summarizing the gradients into

orientation histograms in subregions on a regular grid, we apply an irregular grid with subregions of different sizes. This modification results in a keypoint descriptor that is less sensitive to scale errors. It will be shown that this novel approach has a remarkable impact on the matching performance.

The paper is structured as follows. In the next section we show how scale quantization error can cause a wrong matching result. Section 3 introduces our new irregular grid approach. In section 4 results are presented and the paper ends with a conclusion.

2 The SIFT Descriptor and Scale Quantization Error

The scale invariance of SIFT is achieved by rescaling the input image repeatedly with a Gaussian scale-space kernel. Feature detection is performed on every scale space image. Obviously, computing more images, increases the accuracy of the scale of a given feature and the more characteristic a descriptor of the feature becomes. Unfortunately, the more images are processed the higher the computational cost. Keeping the number of necessary scales small is consequently a desirable design goal.

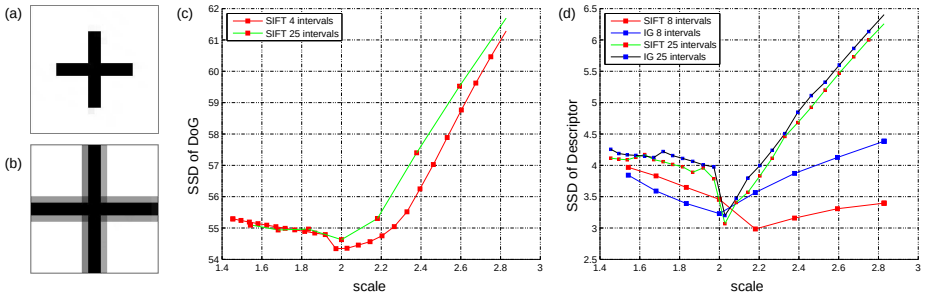

Fig. 1. (a) first image patch, **(b)** second image patch, which is a scaled version of the first patch with scale factor 2.0, **(c)** Sum of Squared Differences (SSD) of the Difference of Gaussian (DoG) for the 2 patches for 4 and 25 scale space intervals, **(d)** SSD of the SIFT descriptors for 4 and 25 scale space intervals, and SSD of the proposed irregular grid (IG) descriptors. For 4 scale space intervals the SIFT descriptor does not show a minimum at a scale of 2.0. In contrast the IG descriptor still has the best SSD at the correct scale.

In many applications, a feature that is detected in one frame has to be redetected in a subsequent image, which has been transformed in various ways. In this paper we focus on scaling between images although other transformations may also be present. Consider the very simple example in figure 1. A cross is shown at two different scales. The difference of Gaussian (DoG), which is the basis of SIFT based feature descriptors, of both images can be computed at different scales. By calculating the sum of squared differences (SSD) of the DoGs

of the two crosses, it is possible to show that by comparing DoGs of different images, the scale factor transforming one cross into the other can be estimated. Figure 1 (c) shows the SSD of DoGs of the two crosses, where scale space is sampled a different number of times. Both plots have their minimum at the scale closest to the real scale. Yet, figure 1 (d) shows that the SIFT descriptor is unable to detect the feature at the correct scale unless a large number of scales is computed.

3 Irregular Orientation Histogram Binning

The original SIFT descriptor summarizes the gradients around a given feature point into orientation histograms in subregions on a regular grid. Several sampling schemes have been proposed in the literature [5,6]. Three of the most common ones are displayed in figure 2.

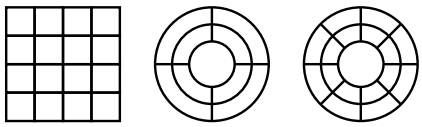

Fig. 2. Left to right: The histogram sampling strategies of the SIFT descriptor, the log polar grid, and the GLOH descriptor.

Consider the matching problem shown in figure 3. The first two images in the top row can be transformed into each other by a single scale s. The matching algorithm processes only a small number of frames to speed up the computation. Assume that the resulting quantization of scale space is so coarse that both descriptors fall into the same interval and, therefore, the scale difference is not compensated. Since the images are related by a scale only, we can transform the regions for which the statistics of the descriptors are collected back into the original image. The predictable similarity of the two descriptors is directly dependent on the overlap areas of the bins of the descriptor. So the design goal of a descriptor that is robust to scale quantization error should be to maximize the overlap of corresponding bins of descriptors when the support regions differ in scale.

When considering the regular 4×4 binning grid on 8×8 pixels, as proposed by Lowe for the SIFT descriptor, the size of the overlapping region R when one of the descriptors is scaled by s can be calculated. For $s < 0.5$, there is no overlap between the outer bins of the original SIFT bins and inner bins overlap s^2 of the area. In contrast, in the irregular grid description, all bins overlap s^2. However, this case is irrelevant in practice because adjacent scales are never separated by more than a factor of 0.5. In fact, normally the scale is closer to 0.8. More interestingly, for $0.5 \leq s < 1$ overlapping region for an inner, outer, and mixed

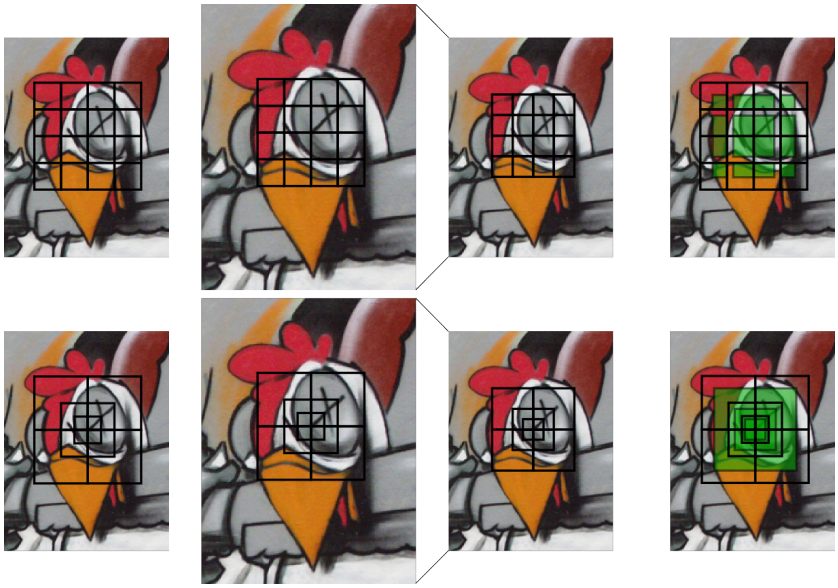

Fig. 3. On the left two images of the same scene are shown. In both images the same feature is detected and the support regions of the bins of the SIFT descriptor (**top**) and of the proposed irregular grid sampling approach (**bottom**) are shown. In the third column the support region of the scaled image is transformed back into the original image. In the right column the regions of overlap of corresponding bins of the scaled descriptors are colored green. It is easy to see, that the green areas for the proposed method are larger than those for the classic SIFT approach.

region, as defined by Fig. 4, amounts to

$$R_{\text{inner}} = 4s^2, \tag{1}$$
$$R_{\text{outer}} = (4s-2)^2, \tag{2}$$
$$R_{\text{mixed}} = 2s(4s-2). \tag{3}$$

So, the average overlap per pixel of the scaled regular grid can be computed by

$$R_{\text{SIFT}} = \frac{1}{64}\left(4R_{\text{outer}} + 4R_{\text{inner}} + 8R_{\text{mixed}}\right) = \frac{9}{4}s^2 - \frac{3}{2}s + \frac{1}{4} \tag{4}$$

Closer analysis of these terms reveals that the bins that are farther away from the center of the feature lose their region of overlap the fastest.

So, instead of a regular grid, we propose to use the arrangement of bins shown in figure 5. The increased size and the absence of an inner boundary of regions that are farther away from the center improves the overlap in presence of scale quantization error. In fact, for all squares the region of overlap can be computed

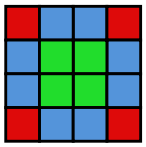

Fig. 4. The inner regions R_{inner} (green) are affected the least by scale quantization error because they reach to the center of the feature. In contrast, corner regions R_{outer} (red) are affected the most. The intermediate regions R_{mixed} (blue) inherit properties from both and are consequently affected moderately.

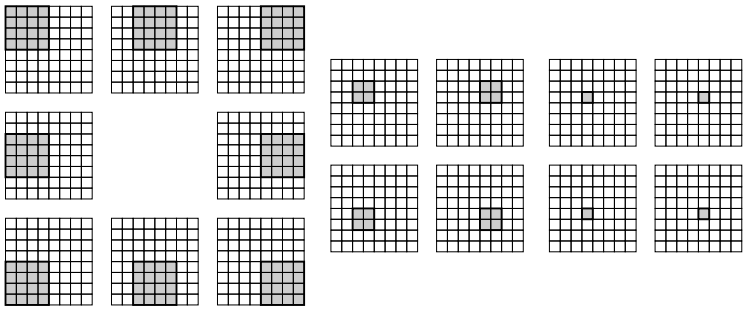

Fig. 5. The proposed sampling strategy does not use a grid or ring structure like previous methods (cf. fig. 2). Instead, all regions extend to the center of the feature. The inner region is consequently sampled several times by different bins. This allows us to drop the Gaussian weighting of the bins used by the SIFT descriptor.

by

$$R_k = (k \cdot s)^2, \tag{5}$$

with $k \in 1, 2, 4$. The average overlap per pixel of the irregular grid is then

$$R_{\text{IG}} = \frac{1}{148}(8R_4 + 4R_2 + 4R_1) = s^2, \tag{6}$$

and thus

$$R_{\text{SIFT}} < R_{\text{IG}}, \quad \text{for } 0.5 \leq s < 1. \tag{7}$$

Other binning schemes proposed in the literature exhibit similarly undesirable overlap progression under scaling. Since the innermost pixels are sampled several times by different bins, an implicit weighting scheme is applied that weights inner regions higher than outer ones. This implicit weighting allows us to drop the Gaussian weighting of the bins suggested by Lowe [5].

In the following section we show that the improved robustness to scale quantization error significantly increases the recall precision compared to the classic SIFT descriptor, although all other parts of the SIFT algorithm are left untouched.

4 Results

In this section a comparison of the proposed irregular grid (IG) sampling method with a number of well known feature descriptors is performed on the image dataset introduced for performance evaluation by Mikolajczyk and Schmid [6][1].

Results are shown for Complex Filters(CF) [7], Gradient Location and Orientation Histograms (GLOH) [6], Steerable Filters (SF) [8], Differential Invariants (DI) [9], Moment invariants (MOM) [10], PCA-SIFT (PCA-SIFT) [11], SIFT (SIFT) [5], Spin images (SPIN) [12], and Cross Correlation (CC). For the results of CF, GLOH, SF, DI, MOM, PCA-SIFT, SIFT, SPIN, and CC we used the code provided by the Visual Geometry Group, University of Oxford[2].

The images we use are compiled in figure 6. The task for all feature descriptors is to find the correct corresponding feature pairs between features detected in the images of the leftmost column and one of the images of the other two columns.

All descriptors work on the same set of features detected by our implementation of the SIFT detector. Since not all details of the original implementation were published by Lowe our algorithm detects slightly different feature point sets. However, the comparison we perform is still fair because all descriptors use the exact same keypoint locations.

Figures 7 and 8 show the points correctly detected by the classic SIFT algorithms on the different input images in blue and the additional correct points detected by the irregular grid (IG) algorithm in red. Also, the recall precision of a number of state-of-the-art algorithms as a function of the total number of matches is given, where recall precision is the ratio between the number of correct matches and number of possible matches

$$\text{recall precision} = \frac{\#\text{correct matches}}{\#\text{possible matches}},$$

and the number of possible matches is defined as the smaller number of feature points detected in either of the input images. The total number of matches N is equal to the number of correct plus the number of false matches.

$$N = \#\text{correct matches} + \#\text{false matches}$$

The total number of matches can be varied by changing the threshold for the maximum allowed distance between two descriptors. The classification into correct matches and false matches is done based on the ground truth transformations that are available for the test images. Our algorithm consistently performs better than the other approaches on all test images.

Please note that in all test scenarios but the second graffiti example (Fig. 6 c(3)) the precision axis is plotted from zero to one. In the second graffiti example, however, the performance of all evaluated descriptors is rather weak because the transformation between the images cannot be approximated very well by rotation and scaling alone. Instead, a strong affine transformation effectively confuses

[1] http://www.robots.ox.ac.uk/∼vgg/research/affine/index.html
[2] http://www.robots.ox.ac.uk/∼vgg/research/affine/descriptors.html

the merely rotation and scale invariant descriptors. The proposed irregular grid descriptor still outperforms the other investigated methods.

5 Conclusion

In this paper a modification of the binning method employed by the classic SIFT descriptor is proposed, which significantly improves the recall precision of the algorithm. The main observation leading to the improved approach is that the overlap of ring or grid based binning schemes diminishes quickly in the presence of scale quantization error. By working with many scales of the input images, this effect can be countered effectively but this is computationally expensive. The presented approach, however, improves the robustness to scale quantization errors at no additional computational cost. We show that recall precision of the modified descriptor consistently outperforms SIFT and several other state-of-the-art descriptors on a standard dataset.

References

1. Tuytelaars, T., Van Gool, L.: Wide baseline stereo matching based on local, affinely invariant regions. In: Proc. British Machine Vision Conference. (2000) 412–425
2. Thormählen, T., Hasler, N., Wand, M., Seidel, H.P.: Merging of feature tracks for camera motion estimation from video. In: 5th European Conference on Visual Media Production (CVMP 2008), London, UK (2008)
3. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. International Journal of Computer Vision **74**(1) (2007) 59–73
4. Sivic, J., Zisserman, A.: Video data mining using configurations of viewpoint invariant regions. Computer Vision and Pattern Recognition, IEEE Computer Society **1** (2004) 488–495
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
6. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Trans. on Pattern Analysis and Machine Intelligence **27**(10) (2005) 1615–1630
7. Baumberg, A.: Reliable feature matching across widely separated views. Computer Vision and Pattern Recognition, IEEE Computer Society **1** (2000) 1774ff.
8. Freeman, W., Adelson, E.: The design and use of steerable filters. Trans. on Pattern Analysis and Machine Intelligence **13**(9) (1991) 891–906
9. Florack, L.M.J., ter Haar Romeny, B., Koenderink, J.J., Viergever, M.A.: General intensity transformations and differential invariants. Journal of Mathematical Imaging and Vision **4** (1994) 171–187
10. Gool, L.J.V., Moons, T., Ungureanu, D.: Affine/ photometric invariants for planar intensity patterns. In: ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume I, London, UK, Springer-Verlag (1996) 642–651
11. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. Computer Vision and Pattern Recognition, IEEE Computer Society **2** (2004) 506–513
12. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using affine-invariant regions. Computer Vision and Pattern Recognition, IEEE Computer Society **2** (2003) 319ff.

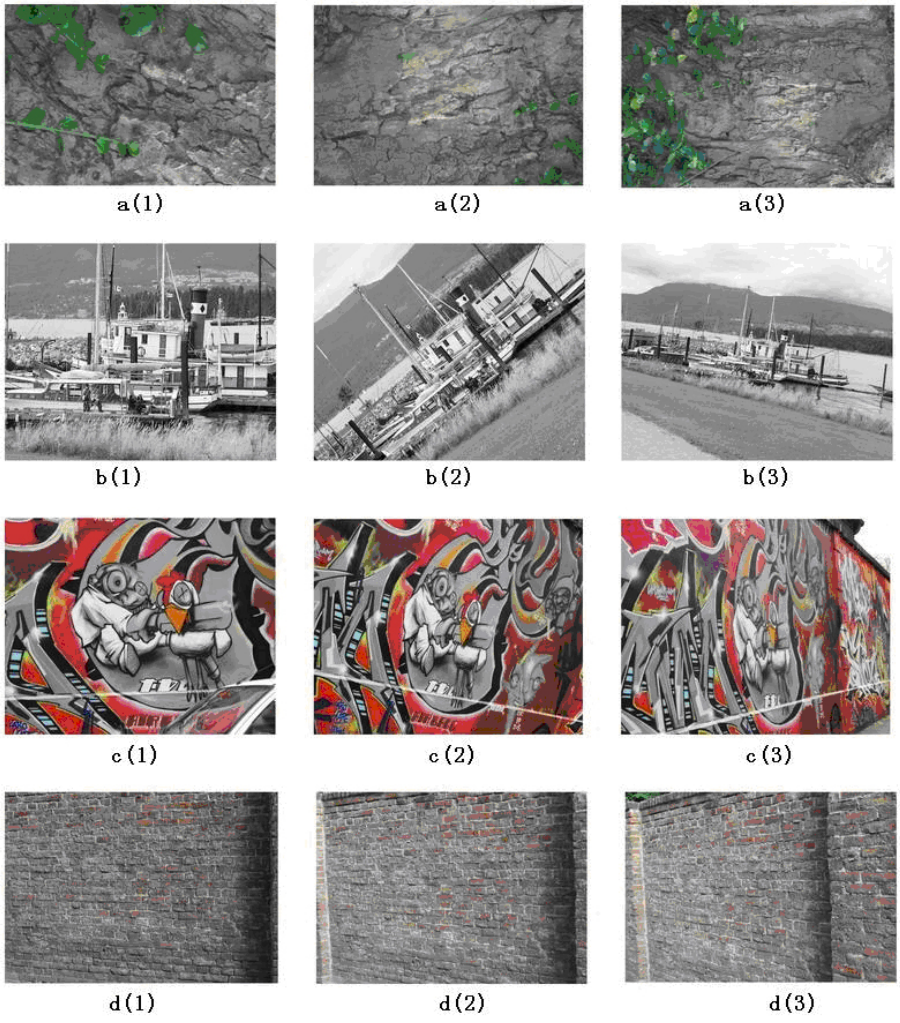

Fig. 6. The input images used to compare the proposed algorithm with state-of-the-art alternatives.

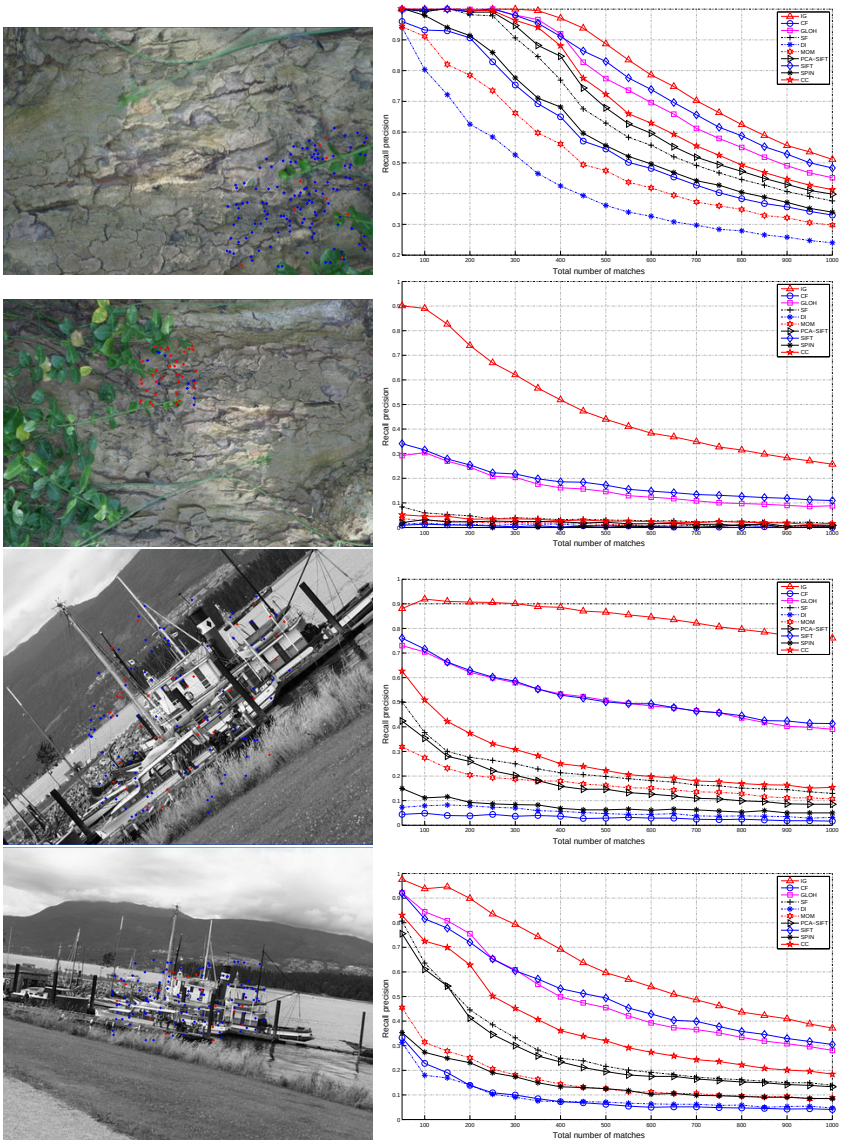

Fig. 7. Left: Blue points are correctly identified by the classic SIFT algorithm and red dots indicate additional correct points found by our method. Results are shown for $N = 200$. **Right**: Recall precision of detected features as a function of the total number of matches N.

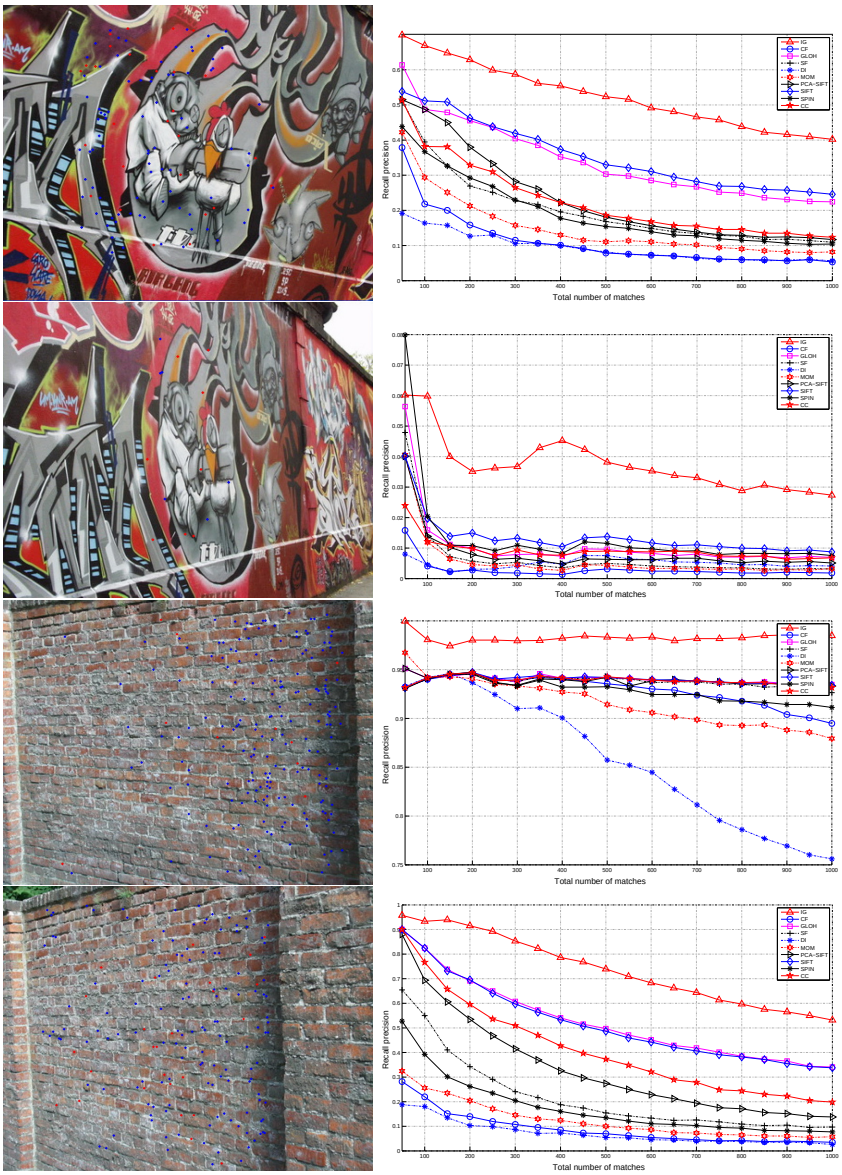

Fig. 8. Left: Blue points are correctly identified by the classic SIFT algorithm and red dots indicate additional correct points found by our method. Results are shown for $N = 200$, except for the second row, where $N = 500$. **Right**: Recall precision of detected features as a function of the total number of matches N.