Feature Points Densification and Refinement

Andrey Bushnevskiy(✉), Lorenzo Sorgi, and Bodo Rosenhahn

Leibniz University of Hannover, Appelstraße 9a, 30167 Hannover, Germany
{andrey.bushnevskiy,bodo.rosenhahn}@tnt.uni-hannover.de

Abstract. A large part of computer vision algorithms and tools rely on feature points as an input data for the future computations. Given multiple views of the same scene, the features, extracted from each of the views can be matched, establishing correspondences between pairs of points and allowing their use in higher-level computer vision applications, such as 3D scene reconstruction, camera pose estimation and many others. Nevertheless, two matching features often do not represent the same physical 3D point in the scene, which may have a negative impact on the accuracy of all the further processing. In this work we suggest a feature refinement technique based on a Harris corner detector, which replaces a set of initially detected feature points with a more accurate and dense set of matching features.

Keywords: Feature points · Densification · Dense · Reconstruction · SIFT · SURF · GFTT · KLT · Harris corners · FREAK

1 Introduction

Feature points extraction is a powerful tool, which has found multiple applications in the field of computer vision. Features are descriptive points, which, being extracted from multiple views of the same scene, are to be matched and further applied in higher-level algorithms, i.e. 3D reconstruction, camera pose estimation, SfM and many others. The specific challenges while working with feature points are improving the performance of the extraction task, minimizing the number of incorrectly identified matches, ensuring localization accuracy of the points in detected matches with respect to the 3D points of the captured scene.

Among the most popular feature point detectors are Harris corner detector [1] and GFTT (Good Features to Track) [2], which, however, do not provide the scale and rotation invariance. Thus, often an additional data structure, called a descriptor, is used for feature points comparison and matching. One of the most well-known descriptor-based feature types is SIFT (Scale Invariant Feature Transform) [3], which is providing invariance to a uniform scale, rotation and partially to affine distortion. The SURF (Speeded up robust features) detector and descriptor based on a fast Hessian detector approximation and a gradient-based descriptor is presented in [4]. In the [5] FREAK (Fast Retina Keypoint) keypoint descriptor inspired by the human retina has been presented, which

S. Battiato et al. (Eds.): ICIAP 2017, Part I, LNCS 10484, pp. 530–538, 2017.
https://doi.org/10.1007/978-3-319-68560-1_47

also provides rotation and scale invariance as well as an advantage in terms of performance. The performance of several types of feature point descriptors has been evaluated under different conditions in [6], confirming the advantages and robustness of SIFT descriptor. The performance of a number of feature detectors and descriptors has also been evaluated in [7] for the task of 3D object recognition. The results of the comparison suggest that SIFT and affine rectified [8] detectors the are the best choice for the task due to their robustness to change of viewpoint as well as changes in lighting and scale.

A new type of scale-invariant feature points is presented in [9]. There the Harris corner detector is combined with SIFT descriptor in order to obtain scale invariance and achieve real-time performance for the tasks of tracking and object recognition by skipping a time consuming scale space analysis. Recent works are applying a deep learning approach to the task of feature extraction. The LIFT (Learned Invariant Feature Transform) [10] presents a deep network architecture trained using a sparse multi-view 3D reconstruction of a scene, which implements three pipeline components, namely feature detection, feature orientation estimation and descriptor extraction.

In this paper we are presenting a novel approach for replacing an initial set of SIFT or other type of feature points with a new and more accurate set of Harris corner matches, extracted from the local neighbourhoods of the matching pairs of the initial set. We test the performance and demonstrate the efficiency of the proposed approach for the tasks of camera pose estimation and sparse point cloud reconstruction.

2 Feature Points Densification and Refinement

Typically, the task of scale-invariant feature extraction is performed on scaled-down versions of original images in order to improve the performance, ensure robustness of the algorithm and maximize the number of correctly detected feature matches [11,12]. Feature points in a correct match, however, often do not represent the same physical 3D point of the object. If the feature point in the first image is considered a reference, the matching feature in the second image may be displaced from a corresponding image point by a few pixels (Fig. 1), which affects the accuracy of the further processing. The number of extracted features may also be significantly reduced for the same reason. Moreover, for the tasks of 3D scene reconstruction and representation, the most descriptive and suitable points are corners, which may be naturally omitted by some of the feature detectors [13].

The approach presented in this paper is aimed at handling these factors by providing a new set of precise corner points, allowing for an accuracy improvement for all the further computer vision applications. The proposed feature densification pipeline is comprised of three steps, namely feature initialization, iterative feature patch warp and tracking of new refined feature points.

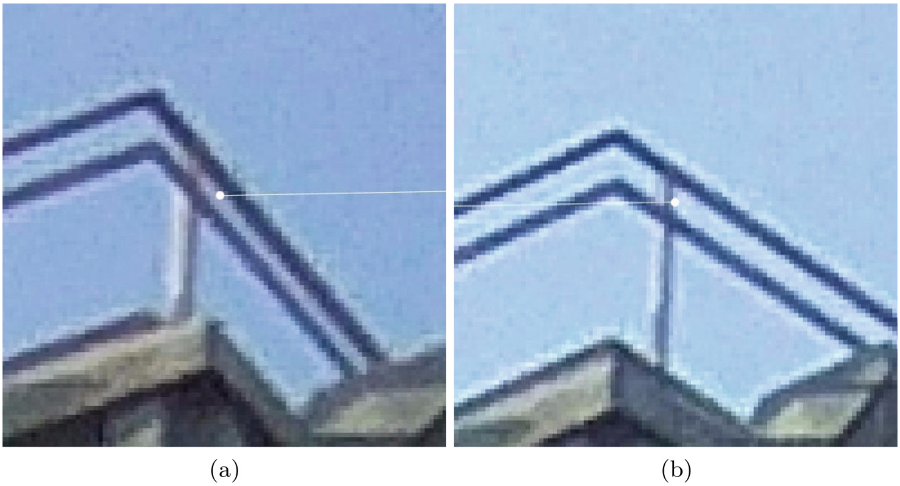


Fig. 1. SIFT features in the original images. Corresponding matching SIFT feature points in the left and right images are having a noticeable displacement.

2.1 Initialization

The proposed algorithm requires an initial set of conventional feature points and matches to be detected in the corresponding pairs of scaled-down images. In this paper we are considering SIFT feature points, however, the approach can also be adapted to the other types of features, such as SURF, FREAK or GFTT.

2.2 Feature Patches Warp

Each feature point depicts image content in its neighborhood, which can be described by an image patch with its center coinciding with the feature point location (Fig. 2(a)). Since two matching features represent the same 3D point of the

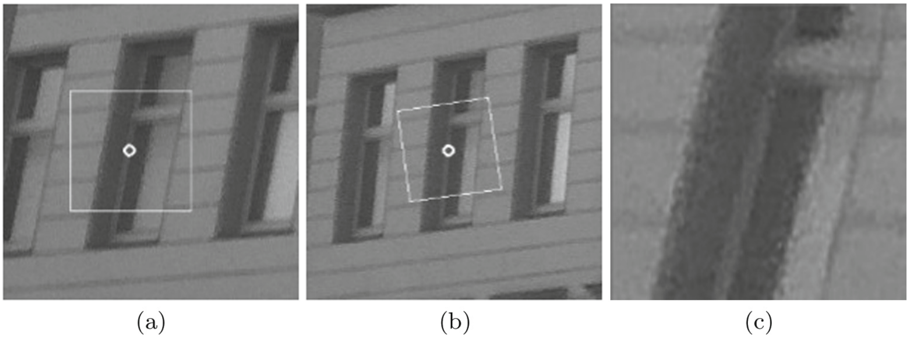


Fig. 2. A patch in the reference image (a) and the corresponding patch in the target image (b), warped using the estimated homography H (c).

captured scene, their corresponding image patches would represent the same area of the scene. Therefore, a new search for matching feature points can be performed locally in corresponding patches of each pair of initial matching features.

Nevertheless, in case of using scale and rotation invariant features (i.e. SIFT, SURF), two image patches have to be transformed before a local feature search can be performed in order to compensate the differences in the scale and orientation of their seed feature points. If one of the images is considered a reference and second a target, for each feature match, it is possible to define a homography, which is relating the reference image patch and the target image:

$$p_t = H \cdot p_{p_r}, \tag{1}$$

where $p_t = (x_t, y_t)$ and $p_{p_r} = (x_{p_r}, y_{p_r})$ are the points in the target image and the reference patch respectively. The size of the reference patch can be defined with respect to the scale of the seed feature point using a user-defined multiplication factor (1.3–2.7 in our experiments). The homography H can be approximated using the positions of two matching feature points together with their orientation, and scale parameters:

$$H = T_2 \cdot S_2 \cdot R_2 \cdot R_1^{-1} \cdot S_1^{-1} \cdot T_1^{-1}, \tag{2}$$

where

$$T_1^{-1} = \begin{bmatrix} 0 & 0 & -x_r + x_{p_{rc}} \\ 0 & 0 & -y_r + y_{p_{rc}} \\ 0 & 0 & 0 \end{bmatrix}, T_2 = \begin{bmatrix} 0 & 0 & -x_t \\ 0 & 0 & -y_t \\ 0 & 0 & 0 \end{bmatrix}, \tag{3}$$

$p_{p_{rc}} = (x_{p_{rc}}, y_{p_{rc}})$ is the top left corner point of the feature patch in the reference image, R_1 and R_2 are the rotation matrices built using orientation angles of the features, S_1 and S_2 are the corresponding feature scale matrices.

Once the homography H is known, the target image (Fig. 2(b)) can be warped and cropped to the target patch (Fig. 2(c)) representing the same part of the scene as the reference, allowing for extraction and tracking of a new feature set.

2.3 Feature Densification and Refinement

The new set of feature points is first extracted from the reference patch using the Harris corner detector [1]. The detected points are then tracked in the transformed target image patch using an iterative Lucas-Kanade tracker [14] (Fig. 3). It is important to mention, that the number of tracked point in the target patch depends on the patch content as well as the size of the reference patch and the quality of the initially detected feature match. Thus, one feature point in the initial set may produce multiple feature points within one patch in a refined set.

The newly extracted and tracked features are then brought back to the reference and target image domains using the homography H:

$$\begin{cases} x_r = x_{p_r} + x_{p_{rc}}, \\ y_r = y_{p_r} + y_{p_{rc}} \end{cases} \tag{4}$$

and

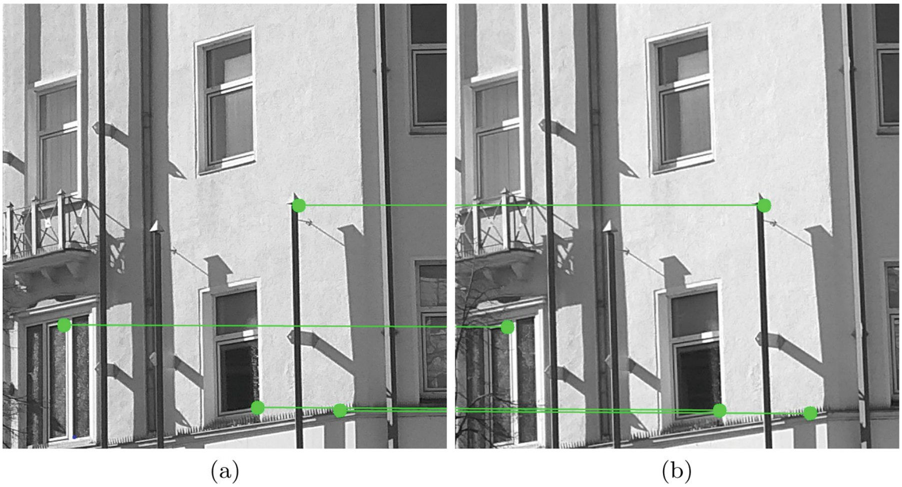


Fig. 3. A reference patch with extracted corners (a) and a target patch with the tracked points (b).

$$\begin{cases} x_t = \dfrac{h_{00} \cdot x_{p_t} + h_{01} \cdot y_{p_t} + h_{02}}{h_{20} \cdot x_{p_t} + h_{21} \cdot y_{p_t} + h_{22}}, \\ y_t = \dfrac{h_{10} \cdot x_{p_t} + h_{11} \cdot y_{p_t} + h_{12}}{h_{20} \cdot x_{p_t} + h_{21} \cdot y_{p_t} + h_{22}} \end{cases}, \qquad (5)$$

where (x_r, y_r) are the coordinates of the new feature, extracted from the reference patch, in the reference image and (x_t, y_t) are the coordinates of the matching feature in the target image.

The points extracted from the reference image and their matches tracked in the target image are then added to the new feature set and the next match from the initial set is processed (Fig. 4).

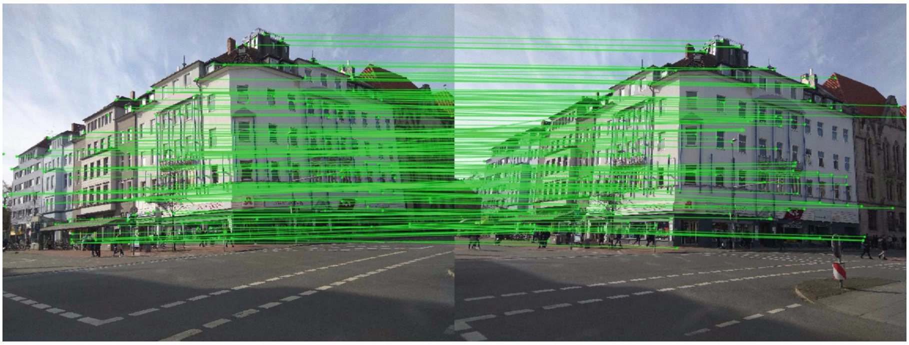


Fig. 4. A set of refined feature points matches.

3 Results

In order to provide a quantitative evaluation of the proposed approach, we have created a dataset comprised of 30 stereo image pairs taken at a 12 MP resolution, using a calibrated camera of a mobile device.

For each image pair, we have performed the tasks of SIFT and SURF features extraction and matching using the scaled-down versions of the original images with the maximum image width of 1024 px. The feature points of this initial set then have been refined using the proposed method (Fig. 4). Each of two feature sets has been used for estimation of the camera poses using the approach described in [15] and triangulation of a sparse point cloud. The set of 3D points has been reprojected back on the images using the corresponding camera poses and camera model parameters. The error then has been evaluated as a pairwise Euclidean distance in pixels between the originally detected feature points and the backprojected point cloud (Fig. 6).

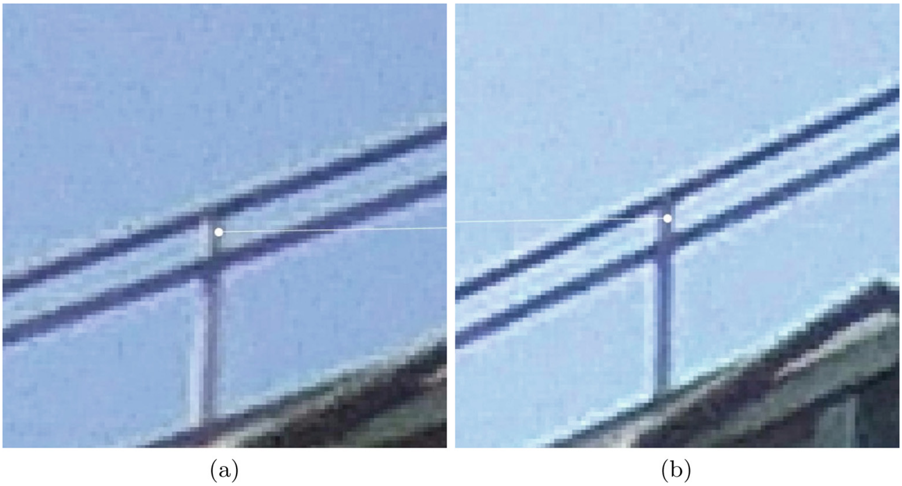
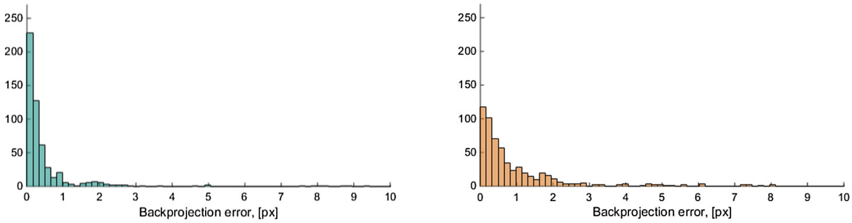


Fig. 5. A feature match from the refined set. Localization error is practically non-existent.

The results, presented in the Table 1, show that the proposed method allows for a significant increase in localization accuracy of the detected feature matches (Fig. 5). This accuracy improvement allows for a more precise estimation of the camera poses as well as a point cloud triangulation. A sample dataset image and the triangulated sparse point cloud, estimated using a refined feature set are shown in Fig. 7.

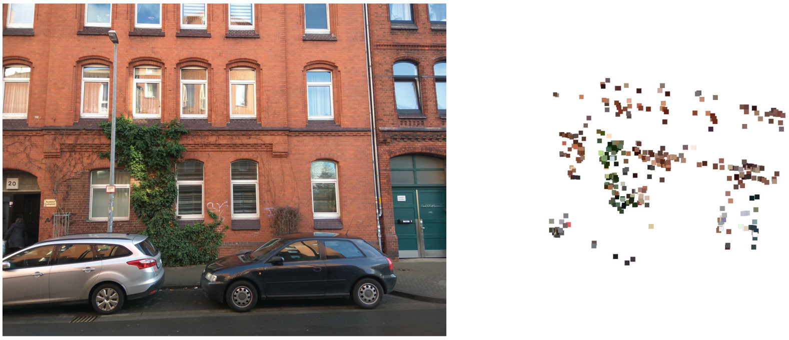
Table 1. Evaluation results. Number of detected feature matches and the 80th percentile of the backprojection error histogram for the initial and refined feature sets.

Type	Number of features	Backproj. error, [px]
Dataset 1		
SIFT	1214	0.60
Refined (SIFT)	1264	0.25
SURF	549	0.84
Refined (SURF)	1109	0.22
Dataset 2		
SIFT	702	1.84
Refined (SIFT)	675	1.06
SURF	215	3.57
Refined (SURF)	384	1.09
Dataset 3		
SIFT	290	0.84
Refined (SIFT)	256	0.67
SURF	65	2.36
Refined (SURF)	102	1.57
Dataset 4		
SIFT	319	2.23
Refined (SIFT)	305	1.02
SURF	103	4.59
Refined (SURF)	223	0.86
Dataset 5		
SIFT	841	0.94
Refined (SIFT)	891	0.54
SURF	178	2.20
Refined (SURF)	492	0.26
Dataset 6		
SIFT	1065	1.05
Refined (SIFT)	1051	0.92
SURF	212	2.42
Refined (SURF)	441	1.08



(a) (b)

Fig. 6. Error histogram for the refined (a) and the initial SIFT features (b).



(a)

(b)

Fig. 7. A reference image and the corresponding sample sparse 3D reconstruction using the feature points from the refined set.

4 Conclusions

The paper presents a new approach for refinement of an initial set of SIFT, SURF or other types feature points. The initial set of matching features is replaced by a new set, obtained by performing a search for Harris corners in the corresponding patches, representing neighborhoods of the original feature points. In contrast to the original one, the new set features an improved localization accuracy as well as a smaller number of incorrectly identified matches. These two factors combined allow for a significant accuracy improvement for the computer vision applications, which are using feature points as an input.

The experimental results prove the efficiency of the proposed approach and demonstrate an accuracy improvement for the tasks of camera pose estimation and a 3D point cloud triangulation using a refined set of matching feature points.

References

1. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference, pp. 147–151 (1988)
2. Shi, J., Tomasi, C.: Good features to track. Technical report, Ithaca, NY, USA (1993)
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
4. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
5. Ortiz, R.: FREAK: fast retina keypoint. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR 2012, Washington, DC, pp. 510–517. IEEE Computer Society (2012)
6. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
7. Moreels, P., Perona, P.: Evaluation of features detectors and descriptors based on 3D objects. Int. J. Comput. Vis. **73**(3), 263–284 (2007)

8. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 128–142. Springer, Heidelberg (2002). doi:10.1007/3-540-47969-4_9
9. Azad, P., Asfour, T., Dillmann, R.: Combining Harris interest points and the SIFT descriptor for fast scale-invariant object recognition. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 11–15 October 2009, St. Louis, MO, USA, pp. 4275–4280 (2009)
10. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: learned invariant feature transform. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 467–483. Springer, Cham (2016). doi:10.1007/978-3-319-46466-4_28
11. Aly, M.: Face recognition using SIFT features (2006)
12. Yoshioka, M., Maeda, Y., Omatu, S.: Criterion for optimal image resolution using SIFT. Artif. Life Robot. **14**(1), 24–28 (2009)
13. Peng, K., Chen, X., Zhou, D., Liu, Y.: 3D reconstruction based on SIFT and Harris feature points. In: IEEE International Conference on Robotics and Biomimetics, pp. 960–964 (2009)
14. Bouguet, J.Y.: Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm (2000)
15. Sorgi, L., Bushnevskiy, A.: Two view geometry estimation by determinant minimization. In: Magnenat-Thalmann, N., Richard, P., Linsen, L., Telea, A., Battiato, S., Imai, F.H., Braz, J. (eds.) VISIGRApp, vol. 3, pp. 592–596. SciTePress (2016)