Distinctive Image Features from Scale-Invariant Keypoints

Summary written by Mahban Gholijafari
February 4, 2025

Summary: This paper addresses the problem of extracting distinctive invariant features to reliably match different views of an object or scene, and using them for object recognition. The extracted features are shown to be give robust matching for varying range of affine distortion, change in viewpoint and illumination, and addition of noise.

Related work: The paper states that image matching started at 1981 with corner detection with the application on stereo matching and short-range motion tracking, and later for more difficult problems ([7], [5], [10]). Torr (1995)[9] developed an approch for long-range motion matching with geometric constraints for rigid objects moving with an image. Schmid and Mohr used the previous work to extend the application to general image recognition with the power to match features with arbitrary orientation change between two images, occlusion, and clutter. Lowe (1999)[6] fixed the problem of corner detectors being sensitive to images with different sizes. They also introduced a new local descriptor that gave more distinctive features that were less sensitive to viewpoint change. The current paper also recognizes some works addressing stability under scale change [3], extending local features to be invariant to full affine transformation [1], and describing phase-based local features to improve invariance to illumination [2].

Approach: This paper presents stages of computation to generate the set of image features that are suitable for image matching of differing objects and or scene. The extracted features are highly distinctive, invariant to image scaling and rotation, partially invariant to change in illumination and 3D camera viewpoint, and are less likely to be disrupted by occlusion, clutter, or noise. This approach of stage based computation is called Scale Invariant Feature Transform (SIFT). With this approach an image of size 500x500 pixels will give about 2000 stable features. To reduce the cost of extracting these features, the author has used cascade filtering approach. The stages of computation are: scale-space extrema detection, keypoint localization, orientation assignment, keypoint descriptor.

The first stage is to identify locations and scales that can be repeatably assigned under different views of the same object. Stable keypoint locations can be detected using scale-space extrema in the difference-of-Gaussian function, $D(x, y, \sigma)$, convolved with the image. Local maxima and minima of $D(x, y, \sigma)$ can be detected by comparing each sample point to its eight neighbors in the current image and nine neighbors in the scale above and below.

The next step is to perform a detailed fit to the nearby data for location, scale, and ratio of principal curvatures. This paper follows Brown and Lowe 2002 [8] work that uses Taylor expansion of scale-space function, $D(x, y, \sigma)$. The location of the extremum is determined by taking the derivative of this function and setting it to zero. The principal of curvature can be computed from a 2x2 Hessian matrix.

The third step is to assign a consistent orientation to each keypoint based on local image properties. This orientation ensures invariance to image rotation. In order for all the computations to perform in a scale-invariant manner, the scale of the keypoint is used to select the Gaussian smoothed image (L). Also, an orientation histogram covering 360 degree range orientation is formed from the gradient orientations of sample points within a region around the keypoint.

Final step is to compute a descriptor for the local image region that is distinctive and yet invariant to remaining variations like change in illumination. This work follows the approach presented by Edelman, Intrator, and Poggio (1997) [4] that was based upon a model of complex neurons in primary visual cortex. Keypoint descriptors are computed by first sampling the image gradient magnitudes and orientations around the keypoint location by using the scale of it to select the level of Gaussian blur.

Datasets, Experiments and Results: This paper uses two databases one consists of 32 images with about 40,000 keypoints, and the other is larger and has 112 images. The author reports 95% accuracy after addition of $\pm 10\%$ pixel noise. For image recognition implementation, by rejecting matches with a distance ratio greater than 0.8, 90% of the false matches and less than 5% of the correct matches are discarded. The total time to recognize all objects in images are less than 0.3 seconds on a 2GHz Pentium 4 processor.

Strengths: SIFT keypoints presented in this paper are invariant to image rotation and scale and are robust across a substantial range of affine distortion, addition of noise and change in illumination. The large number of keypoints leads to robustness in extracting small objects among clutter. Keypoints extraction is also highly efficient.

Weaknesses: This paper lacks systematic testing for full 3D viewpoint and illumination changes. Also, they only used monochrome intensity image, and did not incorporate texture measures.

Reflections: Other potentioal applications can be view matching for 3D reconstruction, motion tracking and seg-

mentation. Other than systematic testing on data with full 3D viewpoint and illumination, further distinctiveness can be derived from color descriptors that are illumination-invariant. The author also mentions that future research can individually learn features that are suited to recognizing particular object categories. Due to large amount of training data that has been available for different object classes, feature sets will probably contain prior and learned features.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 774–781. IEEE, 2000.

[2] G. Carneiro and A. D. Jepson. Phase-based local features. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7*, pages 282–296. Springer, 2002.

[3] J. L. Crowley and A. C. Parker. A representation for shape based on peaks and ridges in the difference of low-pass transform. *IEEE transactions on pattern analysis and machine intelligence*, (2):156–170, 1984.

[4] S. Edelman, N. Intrator, and T. Poggio. Complex cells and object recognition. 1997.

[5] C. Harris, M. Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[6] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[7] H. P. Moravec. Rover visual obstacle avoidance. In *IJCAI*, volume 81, pages 785–790, 1981.

[8] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *IEEE/RSJ international conference on intelligent robots and systems*, volume 1, pages 226–231. IEEE, 2002.

[9] P. Torr. Motion segmentation and outlier detection phd thesis. *University of Oxford*, 1995.

[10] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial intelligence*, 78(1-2):87–119, 1995.