Distinctive Image Features from Scale-Invariant Keypoints

Summary written by Nicholas Hinke
February 08, 2022

Summary: This paper proposes a novel method of extracting useful identifying features from an image to address a variety of challenges in image feature recognition. Denoted as *Scale-Invariant Feature Transform* (SIFT) descriptors, the author primarily sought to tackle the problems of variance to scale and rotation found among other feature descriptors of the time. Inspired by the then-recent work on stable image extrema [7, 9], affine invariant features [1, 2, 8], image contour grouping [10, 11], and even corner detection-based features [3, 5, 12], the author claims to have extended portions of much of this work in order to develop a computationally-efficient four-stage pipeline for extracting SIFT features from an image. Moreover, this paper studies a variety of potential applications for this class of features, most notably including object recognition via clusters of three or more descriptors [6].

Approach: As noted above, the author implemented a four-stage pipeline in order to extract SIFT features from an image. In sequence, the stages are: 1) scale-space extrema detection, 2) keypoint localization, 3) orientation assignment, and 4) keypoint descriptor representation [6]. For a given keypoint, the output of this process is four parameters–2D keypoint location, scale, and orientation–that yield a locally scale-invariant coordinate frame relative to the feature. Additionally, the author considered runtime efficiency during each stage of the pipeline, and took several steps to substantially improve the algorithm's efficiency [6].

During the first stage of scale-space extrema detection, the goal is to identify any points of interest within the image which are likely to be invariant to scale and rotation. This is done by first convolving the image with a number of "difference-of-Gaussian" functions that vary in scale by a constant factor in order to efficiently approximate the scale-normalized Laplacian of a Gaussian–thus more closely resembling "true scale invariance" [4, 6]. Then, a pixel that represents an extremum within the image–which consequently is a point of interest–can be identified by comparing its value to its 8 neighbors in its current scale as well as to its 18 other neighbors in the scales above and below [6].

The second stage of keypoint localization attempts to determine the location and scale of each keypoint by fitting a model of the scale-space sample function $D(x, y, \sigma)$ at each extremum [2, 6]. Additionally, this model is useful for rejecting keypoints which are unstable due to their relatively low contrast. Also regarding stability, the author attempts to reject strong edge responses during this stage due to their poor robustness to noise; this is done by approximating the principal curvatures of the edge using an estimate of the Hessian [3, 6].

The third stage of orientation assignment is done by assembling a histogram of local image gradient orientations at the same scale as each respective keypoint. If there are multiple dominant peaks within the histogram, additional keypoints will be constructed with the same location and scale but with varying orientations [6].

Finally, the fourth stage of keypoint descriptor representation attempts to combine the four previously computed parameters of each keypoint into a unified representation that is invariant to distortion and variance in illumination. Several important steps are taken to accomplish this, including rotating the descriptor and gradient orientations relative to each keypoint to achieve orientation invariance, as well as modifying each feature vector–by both normalizing it and applying small changes to its brightness and contrast–to achieve illumination invariance [6].

Strengths: Most importantly, SIFT features as proposed by this paper demonstrate such power in their scale and rotation invariance that they are still widely-used today. Notably, the features are not only highly distinctive, but also well-localized in both the spatial and frequency domains [6]. Additionally, the author wisely considered several other relevant factors, including keypoint stability, runtime efficiency, and even invariance to occlusions and illumination when designing these features. Finally, the author also provides a brief survey of further applications of these descriptors beyond just object recognition in images [6].

Weaknesses: Although considered throughout each stage of the descriptor pipeline, it is unclear from the paper how the runtime efficiency would scale to more modern high-resolution images (*e.g.* '4K' images with over 8 million pixels). Additionally, while the author briefly highlights some of the effects of hyperparameter choice, it would be interesting to see a more exhaustive survey of how each parameter was chosen and where further room for improvement remains. Finally, as mentioned by the author, it would be worthwhile to study methods to improve the affine invariance of the descriptors [6].

Reflections: As briefly mentioned in the paper, 3D object pose estimation is possible via clusters of three or more four-parameter keypoints. It would be valuable to attempt to extend this approach in order to provide a better estimate of the full 6 DoF pose. Additionally, all of the images used

in the paper were monochromatic, so it may prove beneficial to study whether or not performance can be improved by using information from multiple color channels [6]. Finally, it would be interesting to study the runtime efficiency of higher-resolution images on the modern GPU's of today.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 774–781, 2000.

[2] M. Brown and D. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference*, volume 13, pages 656–665, 01 2002.

[3] C. G. Harris and M. J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[4] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21:224–270, 09 1994.

[5] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. British Machine Vision Computing 2002.

[8] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *European Conference on Computer Vision (ECCV)*, pages 128–142, Copenhagen, Denmark, 2002.

[9] K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. *Proceedings of the British Machine Vision Conference*, 09 2003.

[10] R. C. Nelson and A. Selinger. Large-scale tests of a keyed, appearance-based 3-d object recognition system. *Vision Research*, 38(15):2469–2488, 1998.

[11] A. Pope and D. Lowe. Probabilistic models of appearance for 3-d object recognition. *International Journal of Computer Vision*, 40(2):149–167, 2000.

[12] 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):87–119, 1995. Special Volume on Computer Vision.