Performance Evaluation of SIFT-Based Descriptors for Object Recognition

Yuehua Tao, Marjorie Skubic, Tony Han, Youming Xia, and Xiaoxiao Chi

***Abstract*—Object recognition has become one of the most active research topics in computer vision in recent years. The set of features extracted from the training image is critical for good object recognition performance. The Scale Invariant Feature Transform (SIFT) was proposed by David Lowe in 1999; the SIFT features are local and effective for object recognition. In this paper we conducted a survey of recent related work on the SIFT descriptor, analyzed the evaluation criteria for object recognition, and analyzed the performance of the SIFT descriptor and extended SIFT descriptors based on common properties and evaluation criterion. The paper documents improvement strategies and trends of the SIFT descriptor and proposed extensions.**

***Index Terms*—feature extraction, object recognition, performance evaluation, SIFT descriptor**

I. Introduction

Object recognition is the task of finding a given object in an image or video sequence. This task represents a significant challenge for computer vision systems. In general, the processing of object recognition has the following stages: feature extraction and feature matching. In computer vision, the Scale-Invariant Feature Transform (SIFT) is an algorithm to identify and characterize local features in images. It allows for correct object identification with low probability of mismatch and is easy to match against a large database of local features. The representation of the image features has a direct impact on the performance of an object recognition system. Thus, the characterization and evaluation of SIFT's performance is important to advance the research on object recognition.

In this paper we survey the recent work done to develop, improve, and evaluate SIFT features. Included in the discussion are an overview of the SIFT descriptor computation (Section II), the development and improvement of SIFT descriptors (Section III), the performance evaluation criterion of SIFT descriptors (Section IV), and conclusions about trends for SIFT descriptors in object recognition (Section V).

Manuscript received December 28, 2009. This work was supported in part by the Key Laboratory construction projects in Yunnan Province, China and the Key Discipline Construction Project in Yunnan Normal University.

Yuehua Tao is a Professor of the College of Computer Science and Information Technology, Yunnan Normal University, Kunming 650092, China (corresponding author, phone: 86-871-5515320; fax: 86-871-5516278; e-mail: tyuehua@ 21cn.com).

Marjorie Skubic is a Professor of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65201, USA (e-mail: skubicm@missouri.edu).

Tony Han is an Assistant Professor of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65201, USA (e-mail: Hantx@missouri.edu).

Youming Xia is with the College of Computer Science and Information Technology, Yunnan Normal University, Kunming 650092, China (e-mail: xyouming@21cn.com)

Xiaoxiao Chi is with the School of Engineering, Deakin University, Melbourne, 3125, Australia (e-mail: xchi@deakin.edu.au)

II. Overview of the SIFT descriptor

The SIFT algorithm is a local feature extraction algorithm, which finds extrema points in scale space, and extracts a position, scale, rotation invariant feature vector for each extrema point. The SIFT descriptor was proposed by David Lowe in 1999 [1], further developed in 2004 [2], and has been the topic of additional improvements in recent years. In order to increase SIFT's power and enhance the efficiency of object recognition, extensions of the SIFT descriptor have been proposed by researchers. The improvement techniques have included different histograms and different region shapes than used for the standard SIFT, and the use of dimensionality reduction methods to reduce the dimension of the standard SIFT descriptor.

The SIFT algorithm finds extrema points in scale space, and extracts position, scale, rotation invariant feature vectors. The major stages of computation of SIFT descriptors are divided into four major stages [2]: (1) scale-space extrema detection (i.e., identifying keypoints); (2) keypoint localization; (3) orientation assignment; and (4) keypoint descriptor computation. These stages are used to produce the set of image features. The sections below provide details for these stages.

A. Scale-Space Extrema Detection

The first stage of calculation is to search over all scales and image locations. The difference-of-Gaussian function is used to detect stable keypoint locations in scale space. This stage attempts to find those locations and scales that are identifiable from different views of the same object. This can be efficiently achieved by using a scale space function. Furthermore, it has been shown under reasonable assumptions that it must be based on a Gaussian function. The scale space of a two-dimensional image is defined by equation (1) as shown below:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

where * is the convolution operator, $G(x, y, \sigma)$ is a variable-scale Gaussian, and $I(x, y)$ is the input image. The parameter σ is the scale of the keypoint and is also the standard deviation of the Gaussian function, equation (2).

ISBN: 978-988-18210-4-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2)$$

The difference of Gaussians function, *D(x, y, σ)*, is used to detect stable keypoint locations in scale space; *D(x, y, σ)* is computed by using the difference between two images, one with scale *k* times the other. Then, *D(x, y, σ)* is given by equation (3).

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y.k\sigma) - G(x, y, \sigma)) * I(x, y) \quad (3) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned}$$

To detect the local maxima and minima of *D(x, y, σ)*, each point is compared with its 8 neighbors at the same scale, and its 9 neighbors up and down one scale. If this value is the minimum or maximum of all these points then this point is an extrema. The extrema is used as a SIFT keypoint.

B. Keypoint Localization

In order to enhance the stability of the follow-up image feature matching and increase the algorithm's anti-noise ability, we need to remove the low-contrast and unstable keypoints. This stage attempts to eliminate these unstable keypoints from the final list of keypoints by finding those that have low contrast or are poorly localized on an edge. This may be achieved by calculating the Laplacian value for each keypoint found in stage one. The location of extremum, *z*, is given by equation (4).

$$z = -\frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x} \quad (4)$$

C. Orientation Assignment

This step aims to assign a consistent orientation to the keypoints based on local image properties. The keypoint descriptor can then be represented relative to this orientation, achieving invariance to rotation. The gradient magnitude, *m,* and orientation, *μ,* of *(x, y)* are given in equations (5) and (6) below.

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (5)$$

$$\mu(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y))) \quad (6)$$

D. The Keypoint Descriptor

The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

The local gradient data, used above, is also used to create keypoint descriptors. The gradient information is rotated to line up with the orientation of the keypoint and then weighted by a Gaussian with a variance of 1.5 * the keypoint scale. These data are then used to create a set of histograms over a window centered on the keypoint.

Keypoint descriptors typically use a set of 16 histograms, aligned in a 4×4 grid, each with 8 orientation bins, one for each of the main compass directions and one for each of the mid-points of these directions. This process results in a feature vector containing 128 elements.

III. Development and Improvement of the SIFT Descriptor

In order to increase SIFT's power, two main directions have been proposed to provide improvement in object recognition. The first uses grayscale images (as in the original SIFT algorithm) and computes SIFT descriptors with different histograms or different region shapes, or aims to reduce the dimensionality. The second uses images in color space (such as HSV or RGB).

To increase illumination invariance and discriminative power, color descriptors have been proposed by a number of researchers [3][4][5][6][7][8]. Koen et al. studied the invariance properties and the distinctiveness of color descriptors in a structured way [9].

PCA-SIFT [10] and GLOH [11] are variants of SIFT that operate with grayscale images. Yan and Sukthankar [10] applied Principal Components Analysis (PCA) to the normalized gradient patch. Mikolajczyk and Schmid proposed the Gradient location-orientation histogram (GLOH) [11] which is designed to increase the robustness and distinctiveness of the keypoint descriptor; this extension changes the location grid and uses PCA to reduce the size.

To obtain a complementary representation and local appearance of normalized patches, Svetlana Lazebnik et al. have developed an additional rotation-invariant descriptor that generalizes Lowe's SIFT, called RIFT (Rotation Invariant Feature Transform) descriptors [12]. RIFT is a rotation-invariant generalization of SIFT.

Other improvements have been proposed to speed up the SIFT computation. Grabner et al. proposed a considerably faster approximation of the well-known SIFT method in 2006 [13], named FA SIFT; it can speed-up the SIFT computation by at least a factor of eight compared to the binaries provided by Lowe. Se et al. implemented SIFT on a Field Programmable Gate Array (FPGA) and improved its speed by an order of magnitude [14].

IV. Performance Evaluation of SIFT Descriptors for Object Recognition

The performance of an object recognition system depends mainly on two compositions: a suitable representation of the image and a powerful image matching and recognition algorithm. In the following sections, we discuss evaluation criterion for object recognition and common properties of local descriptors. On this basis, we conduct an in-depth analysis and discussion of the common properties and recognition performance evaluation methods of the SIFT descriptor and extended SIFT descriptors.

A. Common Properties of SIFT Descriptors

Based on a review of recent literature, we propose the following properties as being important for a good local feature: (1) Must be highly distinctive, robust to changes on viewing conditions as well as to errors of the detector; a good feature should allow for correct object identification with low probability of mismatch; (2) Should be easy to extract; (3)

ISBN: 978-988-18210-4-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Invariance; a good local feature should be tolerant to image noise, changes in illumination, uniform scaling, rotation, and minor changes in viewing direction; (4) Should be easy to match against a potentially large database of local features.

SIFT descriptors have the following properties. SIFT features are local features and are mostly invariant to image translation, rotation, scale reduction and amplification, brightness changes, occlusion and noise, and are stable to visual changes and affine transformation to a certain extent. SIFT features have high distinctiveness and abundant information, and provide fast and exact matching in a large feature database. The computation of SIFT features is relatively fast and optimized; thus, the SIFT matching algorithm may meet the need for real-time operation. Extensibility is strong and may combine conveniently with other feature vectors. In addition, the dimensionality of the descriptor is also very important, because it heavily influences the complexity of the matching process (at runtime) and the memory requirements for storing the descriptors.

B. Performance Evaluation Criterion for Object Recognition

When performing experiments over multiple object classes, the average precision of the individual classes can be aggregated. This aggregation is called the mean average precision (MAP). MAP is calculated by taking the mean of the average precisions. Note that MAP depends on the dataset used; scores of different datasets are not easily comparable.

The performance is also measured by the repeatability rate, that is, the percentage of points simultaneously present in two images. In recent years, recall and precision [10][17], average precision [9], repeatability rate [14], and the ROC curve [11][13][15][16] have become popular evaluation metrics for object recognition.

Recall and precision are based on the number of correct and false matches between two images. Recall is the number of correctly matched regions with respect to the number of corresponding regions between two images of the same scene. Recall is defined as in equation (7).

$$recall = \frac{number\ of\ correct\ positive}{total\ number\ of\ positives} \quad (7)$$

Precision is defined according to equation (8).

$$precision = \frac{number\ of\ correct\ positives}{number\ of\ correct\ positives + number\ of\ false\ positives} \quad (8)$$

The metric *1-precision* is defined as in equation (9).

$$1 - precision = \frac{number\ of\ false\ positives}{total\ number\ of\ matches(correct\ or\ false)} \quad (9)$$

Under normal circumstances, the metrics of recall, precision and 1-precision can be used for object recognition system evaluation.

Average precision is a single-valued measure that is proportional to the area under a precision-recall curve. This value is the average of the precision over all test runs judged relevant. Let $\rho^k = \{l_1, l_2, \cdots, l_k\}$ be the ranked list of items from test set A. At any given rank k, let $|R \cap \rho^k|$ be the number of relevant shots in the top k of ρ, where R is the set of relevant shots and $|X|$ is the size of set X. Average precision, AP, is then defined according to equation (10).

$$AP(\rho) = \frac{1}{|R|}\sum_{k=1}^{|A|}\frac{|R \cap \rho^k|}{k}\Psi(l_k) \quad (10)$$

with the indicator function $\Psi(k) = 1$ if $l_k \in R$ and 0 otherwise, and $|A|$ is the size of the answer set, e.g. the number of items present in the ranking points simultaneously present in two images. The higher the repeatability rate between two images, the more keypoints that can potentially be matched and the better the matching. In [2], Lowe used repeatability to assess performance of the SIFT descriptor.

The Receiver Operating Characteristics (ROC) curve is a plot of the true positive rate vs. the false positive rate [15]. These curves also show ROC *equal error rates* (false positives = true positives) where

$$true\ positive\ rate = \frac{number\ of\ correct\ positives}{total\ number\ of\ positives\ in\ data\ set} \quad (11)$$

$$false\ positive\ rate = \frac{number\ of\ false\ positives}{total\ number\ of\ negatives\ in\ data\ set} \quad (12)$$

C. Performance Evaluation of SIFT Descriptors and Extended SIFT Descriptors

There is prior work in conducting performance evaluation for the SIFT descriptor and extended SIFT descriptors in object recognition. Koen et al. evaluated color descriptors for object and scene recognition [9]. Yan et al. [10] proposed a PCA-SIFT descriptor similar to the SIFT descriptor and used *recall* and $1-precision$ graphs for their evaluation experiments by varying the threshold for each algorithm. Their experiments showed that the PCA-based local descriptors are more distinctive, more robust to image deformations, and more compact than the standard SIFT representation.

Mikolajczyk and Schmid [11] presented a comparative study for several local descriptors. Their evaluation criterion was recall-precision, based on the number of correct matches and the number of false matches obtained for an image pair. The results of their experiments are shown in Table I.

Table I. Results of experiments in [11]

Descriptor	recall	1-precision	#nearest neighbor correct matches
GLOH	0.25	0.52	192
SIFT	0.24	0.56	177
Shape context	0.22	0.59	166
PCA-SIFT	0.19	0.65	139
Moments	0.18	0.67	133
Cross correlation	0.15	0.72	113
Steerable filters	0.12	0.78	90
Spin images	0.09	0.84	64
Differential invariants	0.07	0.87	54
Complex filters	0.06	0.89	44

ISBN: 978-988-18210-4-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table II is a summary of common properties for the grayscale SIFT descriptor and extended grayscale SIFT descriptors. Based on the literature [3][4][5][9], we summarized common properties for the color SIFT descriptor (Table III).

Table II. Properties of the grayscale SIFT descriptor and extended grayscale SIFT descriptors

descriptor	rotation scale invariance	illumination invariance	viewpoint invariance	distinctiveness robustness	dimensionality
SIFT	yes	yes	yes	good	high(128)
PCA-SIFT	yes	yes	yes	good	low (20)
GLOH	yes	yes	yes	enhancement	high(128)
SIFT-FPGA	yes	yes	yes	good	high(128)
RIFT	yes	yes	yes	good	low(32)
FA SIFT	yes	yes	yes	good	high(128)

Table III. Properties of color SIFT descriptors

Descriptor	scale invariant	shift invariant	invariant to light color changes	dimensionality
HSV-SIFT	yes	yes	partial	3×128
HueSIFT	yes	yes	no	3×128
OpponentSIFT	yes	yes	no	3×128
W-SIF	yes	yes	no	3×128
rgSIFT	yes	yes	no	3×128
Transformed color SIFT	yes	yes	yes	3×128

V. Conclusions

The Scale Invariant Feature Transform (SIFT) features are local and based on the appearance of the object at particular interest points. They are invariant to image scale and rotation. They are also robust to changes in illumination, noise, and minor changes in viewpoint. The SIFT descriptor has a wide range of applications and is especially effective for object recognition.

In order to increase SIFT's power, proposed improvement techniques use different histograms and different region shapes compared to the standard SIFT descriptor. To increase illumination invariance and discriminative power, color descriptors have been proposed. Also, dimensionality reduction methods have been used to reduce the complexity of the standard SIFT descriptor. Recall-precision, average precision, repeatability rate, and the ROC curve have become popular evaluation metrics in the literature for object recognition.

References

[1] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features" , Proc. Seventh Int'l Conf. Computer Vision, pp. 1150-1157, 1999.

[2] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", Int'l J. Computer Vision, vol. 2, no. 60, pp. 91-110, 2004.

[3] A.Bosch, A.Zisserman, and X.Muoz, "Scene classification using a hybrid generative/discriminative approach", IEEE Trans. Pattern Analysis and Machine Intell., 30(04):712–727, 2008.

[4] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance" , IEEE Trans. Pattern Analysis and Machine Intell., 23(12):1338–1350, 2001.

[5] J. van deWeijer, T. Gevers, and A. Bagdanov, " Boosting color saliency in image feature detection", IEEE Trans. Pattern Analysis and Machine Intell., 28(1):150–156, 2006.

[6] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W. T Freeman, "Discovering objects and their locations in images", In International Conference on Computer Vision, pp.370–377, Beijing, China, October 2005.

[7] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos" , In International Conference on Computer Vision, volume 2, pp.1470–1477, Nice, France, October 2003.

[8] T. Gevers, J. van de Weijer, and H. Stokman, "Color image processing: methods and applications: color feature detection", chapter 9, pp. 203–226. CRC press, 2006.

[9] Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, "Evaluation of Color Descriptors for Object and Scene Recognition", Proceedings of CVPR. Anchorage, Alaska, USA, June 2008.

[10] Yan Ke, Rahul Sukthankar, " PCA-SIFT: A More Distinctive Representation for Local Image Descriptors", Computer Vision and Pattern Recognition, 2004. CVPR 2004. page(s): II-506- II-513 Vol.2

[11] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", In Proceedings of Computer Vision and Pattern Recognition, pp. 257-264, 2003.

[12] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce, "Semi-Local Affine Parts for Object Recognition", Proceeding of the British Machine Vision Conference, Kingston, UK, September 2004,Vol.2, pp.959-968

[13] M. Grabner, H. Grabner, H. Bischof, "Fast approximated sift", in: Proceedings of the Asian Conference Computer Vision, 1 (2006) 918–927.

[14] Se, S., Ng, H., Jasiobedzki, P., Moyung, T. ,"Vision based modeling and localization for planetary exploration rovers", Proceedings of International Astronautical Congress (2004)

[15] G. Carneiro and A.D. Jepson, "Phase-Based Local Features" , Proc. Seventh European Conf. Computer Vision, pp. 282-296, 2002.

[16] S. Agarwal and D. Roth. "Learning a sparse representation for object detection", In Proceedings of European Conference on Computer Vision, pages 113–130, 2002.

ISBN: 978-988-18210-4-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)