Robust Image Matching with Selected SIFT Descriptors

Jingwei Wang, Seongho Cho and C.-C. Jay Kuo
Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089
jingweiw@usc.edu, seonghoc@usc.edu, cckuo@sipi.usc.edu

***Abstract*—A robust image matching algorithm using a set of selected SIFT descriptors is investigated in this work. We first utilize the color-based segmentation method and the watershed algorithm to separate foreground and background regions in images and then search the corresponding SIFT descriptors along foreground contours. These selected SIFT descriptors can offer more robust and stable image matching results. Furthermore, we reduce the dimension of SIFT descriptors using the skeleton pruning technique that eliminates unessential key points. It is demonstrated by experimental results that the image matching algorithm with the proposed selected SIFT descriptors outperforms the classical SIFT-based image matching algorithm by a significant margin.**

I. Introduction

Image matching is a fundamental problem in computer vision and image processing, which has extensive applications in object recognition, motion tracking, etc. Many matching techniques have been proposed in the last two decades using various image features, such as the SIFT (Scale Invariant Feature Transform) descriptor [1], complex features [2], etc. The performance of an image matching algorithm can be measured by several performance metrics; *e.g.* robustness, accuracy and complexity. In this work, our objective is to develop a robust image matching process using a set of selected SIFT descriptors.

A typical image matching algorithm consists of three steps. First, it selects distinct points of interest in different regions of given images, which are invariant to image distortion/deformation and can be located repeatedly. Second, properties in the neighborhood of points of interest are presented as features, which are also called "descriptors". Third, the descriptors from different images are matched in pair to offer the final matching results. Based on the above discussion, both points of interest and descriptors have significant impact on the performance of image matching algorithms. They should be stable and reliable for detection. Since the SIFT descriptor [1] outperforms other descriptors in terms of stability by utilizing invariant local features, it is widely adopted as image matching solutions. For example, SIFT has been used in metric robot localization after it was first proposed by Lowe in [1]. Moreover, SIFT can be applied in medical imaging, comparative study, etc.

However, the SIFT-based matching algorithm has several shortcomings. First, when two images have a simple foreground object with complex background, the SIFT-based scheme might miss the actual object by spreading points of interest over the entire image (or even in the background region only). In this case, the SIFT descriptor will fail to offer correct matching results. This is especially true when the foreground remains the same while the complex background changes in different images. Second, a large number of SIFT descriptors are often extracted and used in the matching process. As a result, the matching speed is slowed down significantly [3], [4].

To overcome these shortcomings, we propose a novel image matching algorithm that uses selected SIFT descriptors in this work. First, we separate foreground and background regions using the color-based image segmentation and the watershed algorithm. This is needed since foreground objects are usually of interest in most applications. As a result, the proposed algorithm can avoid missing objects of interest and offer a more robust and reliable output than the original SIFT-based matching algorithm. Next, to lower the dimension of the resultant matching problem, we employ the skeleton-pruning technique to eliminate unessential key points. This step helps accelerate the matching process for higher efficiency. Finally, we propose a robust matching process based on selected sample points, which are located in the neighborhood of essential key points, to improve matching accuracy.

The rest of the paper is organized as follows. We first review the SIFT descriptor in Section II. Then, the proposed image matching process using selected SIFT descriptors is described in Section III. Experimental results are shown in Section IV. Finally, concluding remarking and future research directions are given in Section V.

II. Background Review on SIFT Descriptors

The SIFT descriptor was proposed by Lowe [1] to serve as a robust image feature. The computation of the SIFT descriptor consists of the following four major steps.

1) The difference-of-Gaussian (DoG) operation is applied in various scales through the Gaussian pyramid, and local peaks (key points) are extracted accordingly. As a result, the extracted key-points are invariant with respect to image scaling.
2) Unstable key-points are rejected, which is decided by the contrast level and the principal curvature ratio whose values are below some threshold levels.
3) Each stable key-point will be assigned with a dominant orientation, which allows invariance with respect to image rotation.
4) Additional key-points are added when multiple orientations exist within 80 percent threshold of the dominant orientation. In other words, multiple key-point descriptors are located in the same location but with different orientations.

Based on the above design, it is easy to verify that the SIFT descriptor is invariant with respect to image translation, scaling and rotation.

However, the SIFT descriptor has two major limitations.

1) **Complex background interference**
   When an image contains complex background, the SIFT descriptors tend to spread over the entire image rather than being concentrated in the object region. As a result, the actual object can be neglected in the matching process.
2) **High computational complexity**
   Since the number of extracted SIFT descriptors is typically large, the computational cost to match extracted key-points is very high. Lowe [1] proposed a best-bin-first alternative to speed up the matching process at the cost of lower matching accuracy.

It is desirable to develop a robust and efficient image matching algorithm to improve the performance of the existing SIFT technique. To achieve this goal, we describe a new image matching algorithm using a set of selected SIFT descriptors in the next section.

III. Proposed Algorithm

We propose a novel framework for image matching based on selected SIFT descriptors, which consists of the following three modules.

- **Foreground/Background Separation**
  Since the foreground has a higher probability to contain objects of interest, we separate foreground and background regions using image segmentation methods [5]. For simplicity, we use the color-based image segmentation scheme and the watershed algorithm in this work. Then, the extracted foreground contour will be processed furthermore.
- **Skeleton Pruning**
  We employ the skeleton pruning technique [9] to reduce the dimension of descriptors. Besides, unessential key-points on the object contour are eliminated.
- **Selected SIFT-based Image Matching**
  After the above two steps, the classic SIFT-based image matching algorithm [1] is applied to match pruned key points and generate an improved matching result at lower complexity.

The above three steps will be detailed in the following three subsections.

A. Foreground/Background Separation

The image matching process attempts to find the correspondence between two images with the same object(s) by locating common descriptors in both images. It works well when images to be matched have simple background. However, if images have complex background, descriptors extracted from complex background can be confusing. That is, many key points will be generated over the entire image so that objects of interest may become less visible. To give an example, we show two crane images in Fig. 1 (a). Both of them have a white crane as the main object. These images with their extracted SIFT descriptors are shown in Fig. 1 (b), where key points are distributed over the entire images rather than the object area. This makes the image matching task very challenging.

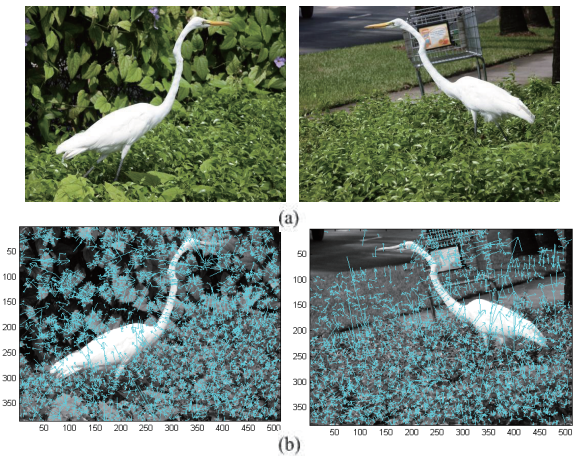

Fig. 1. (a) Two test crane images and (b) their SIFT descriptors.

To address this problem, we perform foreground/background separation so that we can focus on SIFT descriptors in the region of interest (*i.e.* foreground) and discard SIFT descriptors in the complex background. foreground/background separation will affect the overall matching performance in the following cases 1) When background changes, we can still locate foreground objects and match them with similar objects. 2) If the background has complex texture, while foreground's texture is simple, the foreground segmentation will help the detector to focus on foreground objects' matching. Thus helps in improving matching accuracy. We use the image segmentation technique to achieve foreground/background separation. By image segmentation [5], we partition an image into smaller homogeneous regions. There are two common ways to perform image segmentation [5]: 1) contour-based, which detects local changes and 2) region-based, which searches for region similarities. In this work, we adopt two computationally efficient segmentation techniques in cascade: 1) color-based image segmentation [8] (region-based) and 2) the watershed algorithm [6] (contour-based).

First, we use the color segmentation technique proposed in [8] to find the closed region of the foreground. The approach is applied in the RGB color space. It can segment images from two to $K_{\max}$ clusters, where $K_{\max}$ denotes the upper limit on the cluster number. Then, the following validity measure is adopted to determine the optimal cluster number:

$$\text{validity} = \frac{\frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}\|x - z_i\|^2}{\min\left(\|z_i - z_j\|^2\right)} \tag{1}$$

where N is the number of pixels in the image, $2 \le K \le K_{\max}$ is the number of clusters, and Z_i is the cluster center of cluster C_i. The integer value K_{opt} that minimizes the above validity measure is the optimal cluster number.

Segmentation for general images could be challenging. To simplify this task, we focus on images with clear foreground/background separation. That is, the foreground object, or called the region-of-interest (ROI) [7], is the largest closed salient area in the central part of a given image. Besides, it has a good contrast with respect to the background. Under the above assumptions, the foreground object can be separated from background easily via segmentation.

To improve the accuracy of the segmented result, the watershed algorithm [6] is performed in the identified foreground region to refine its boundary contour. The watershed algorithm is applied to the gray-level image, where the gradient of gray level values is exploited for segmentation. Basically, it can segment an image into several homogeneous regions with closed contours, and each of them has similar gray levels. Note that the watershed algorithm may segment an object into multiple regions so that it is often combined with another segmentation method (*e.g.* color segmentation) for better accuracy.

B. Skeleton Pruning

As discussed earlier, another main drawback of the SIFT-based image matching algorithm is the large size of descriptors [3], [4]. To speed up the image matching process, it is desirable to eliminate unessential key points to reduce the descriptor dimension. To match foreground objects, we seek a proper representation of objects, which should provide the shape and topological structure.

The skeleton-based representation has been widely used in image processing and recognition. However, its result may not be stable since the deformation of object's contour may degrade the matching accuracy by introducing erroneous skeleton branches. To overcome this problem, we adopt the skeleton-pruning method [9] which eliminates noise branches yet maintain the topological information. The basic skeleton-pruning procedure is to use the discrete curve evolution to reduce the number of skeleton points. As the curve evolves, smaller convex/concave regions and the associated skeleton points are removed. As a result, only skeleton points aligned with

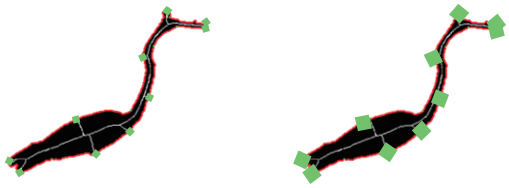

Fig. 2. Illustration of two search window sizes: 10×10 (left) and 20×20 (right)

major corners of the simplified contour remain, and their locations are more robust as compared to the original skeleton points. For verification, we computed locations of pruned skeleton points in 20 objects under different conditions and observed that most of them are confined to a small window of size $W \times W$. The value of W will be studied in Sec. IV.

C. Selected SIFT-based Image Matching

Since the exact location of pruned skeleton points of a similar foreground object may vary, it will affect final matching performance if one conducts the matching of these skeleton points directly. To improve the robustness of the matching algorithm, we uniformly sample n points from the neighborhood of each skeleton point of size $W \times W$. After obtaining the SIFT descriptor at these sampled points, we have SIFT descriptors associated with the n sampled points for matching. There are two parameters required in this process. They are selected based on the following consideration.

- Choice of the number of sampled points - n
  A poor choice of n may lower the final matching performance. On one hand, if n is too large, the dimension of extracted SIFT descriptors will increase, which results in an inefficient matching performance because of the high computational complexity. On the other hand, if n is too small, the matching performance may be unstable since the SIFT descriptors may vary due to the varying location of skeleton points.
- Choice of search window size
  We define the neighborhood of a skeleton point as points along the foreground contour within a window of size $W \times W$, where the orientation of the window is defined by the orientation of the SIFT descriptor at the skeleton point. Clearly, there is a tradeoff between the matching performance and computational complexity. If a window size is too large, it may result in the overlap of two search windows, which in turn may confuse the matching process. In the experimental section, we compare experimental results with different window size and sample points number to show their impact on the matching performance.

In the implementation, we first compute the gradient magnitude and orientation in a region centered at a sampled point. The region is split into $r \times r$ subregions. An orientation histogram for each subregion is formed by accumulating pixels in the subregion, weighted by their gradient magnitudes. Concatenating the histograms from subregions gives a SIFT vector. Furthermore, an explicit scale is determined for each point, which allows the image description vector for that point to be sampled at an equivalent scale in each image. A canonical orientation is determined at each location so that the proposed method is robust to both scaling and rotation.

To match each pair of points, we use the same procedure as given in [1]. Specifically, the best match for each point is found by identifying its nearest neighbor among the points from another group. Also, the nearest neighbor is defined as the point with the minimum $D = \|x_1 - x_2\|_2$, which D is the Euclidean distance of positions x_1 and x_2 of two SIFT descriptors.

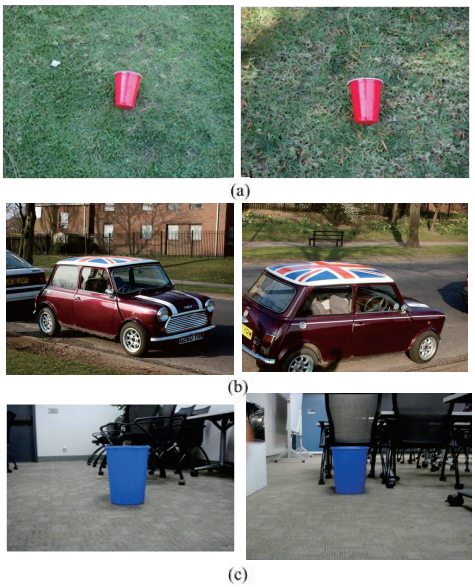

Fig. 3. Illustration of matching examples (a) matching group 2 (b) matching group 3 (c) matching group 4.

IV. Experimental Results

We present experimental results in this section to validate the robustness of the proposed algorithm using selected SIFT descriptors. The experiments were conducted on four pairs of test images as shown in Fig. 1(a) (called group #1) and groups #2, #3, #4 in Fig. 3.

As compared to the traditional SIFT matching process, the proposed foreground/background separation scheme improves the overall matching performance in two aspects. First, if the background has complex texture, the proposed scheme helps the detector focus on the matching of foreground objects as shown in groups #1 and #2. Second, when the background varies, we can still identify foreground objects for the matching task as shown in groups #3 and #4. Thus, the separation of foreground and background is critical to the improvement of matching accuracy.

It is desirable to understand the impact of the number of sampled points in a window and the window size. First, we fix the window size and plot the number of matched points as a function of the number of sampled points in Fig. 4(a). Although more sampled points allows more matching points, there exists a ceiling. Next, we show the number of matched points as a function of the window size parameterized by the number of sampled points in Fig. 4(b). We observed that there is a good range of window sizes, which is between 20 and 30. The matching performance degrades if the window size is too large or too small. Generally speaking, the higher sample density the better performance. However, when a window size is just large enough so that two neighboring windows overlap with each other, the performance reaches the peak.

For the purpose of performance benchmarking, we compare the proposed algorithm with the original SIFT and PCA-SIFT [10]. Here, we set the window size to 20×20 and the number of sampled points to two different values; namely, 10 and 20. As shown in Table I, the original SIFT and the PCA-SIFT algorithms return zero correct

TABLE I
MATCHING RESULTS

Matching Results	group1	group2	group3	group4
Original SIFT	0	0	2	0
Matching Ratio	0/3854	0/5107	2/1187	0/960
PCA-SIFT	0	0	4	0
Matching Ratio	0/3850	2/5008	4/1180	1/974
Proposed(10 samples)	8	6	12	6
Matching Ratio	80%	60%	57%	67%
Proposed(20 samples)	10	9	15	7
Matching Ratio	100%	90%	71%	78%

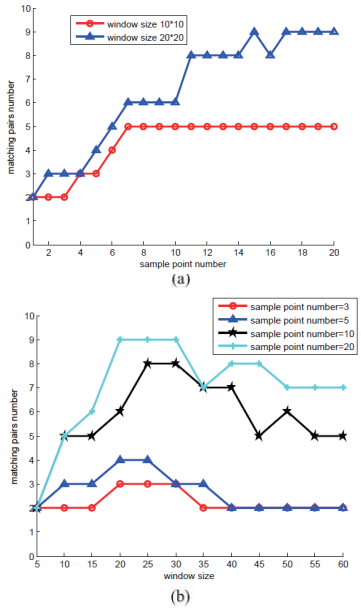

Fig. 4. Matching results of the proposed algorithm as a function of (a) the number of sample points and (b) the window size.

matching pair except for Group 3. In contrast, the proposed algorithm with 10 sampled points can return 8, 6, 12, 6 correct matches, respectively. If the number of sampled points goes to 20, the matching results are even better. The correct matched numbers become 10, 9, 15, 7, respectively. The matching ratio (or the recall rate) for each test image pair is also given in the table. The averaged matching ratios are equal to 66% and 85%, respectively, for 10 and 20 sampled points. The use of a larger number of sampled points offers a better result since it provides more points to match at the cost of higher complexity. The original SIFT and the PCA-SIFT algorithms perform poorly due to the interference of complex background, which is well overcame by the proposed algorithm using selected SIFT descriptors.

The matching results for group #1 with a window of size 20×20 and 10 or 20 sampled points are shown in Fig. 5 (a) and (b), respectively. They share eight common pairs distributed in the head (1), beak (2), neck (2), body's back (1), leg (1) and tail (1) regions. There are two more pairs (one in the neck and the other in the tail), which are missed when the number of sampled points is 10 but recovered by increasing the number of sampled points from 10 to 20.

V. CONCLUSION AND FUTURE WORK

A robust algorithm for image matching using selected SIFT descriptors was proposed in this work. The algorithm achieves better

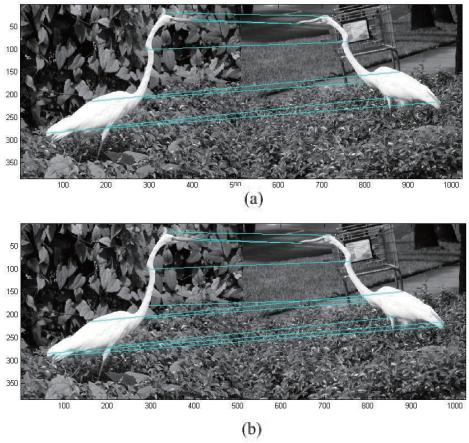

Fig. 5. Comparison of matching results for the proposed algorithm with (a) 10 and (b) 20 sampled points.

efficiency by eliminating unessential key points via foreground contour extraction and skeleton pruning. The robustness and flexibility of the proposed algorithm were demonstrated by experimental results. Currently, the time of the proposed algorithm spent on the matching process is about 10% of that the original SIFT process. However, the proposed algorithm demands image pre-processing operations. As an extension of the current work, we would like to examine ways to speed up these image pre-processing steps to reduce the overall complexity. Also, the proposed algorithm demands images that have clear foreground/background separation, it is desirable to have an advanced segmentation algorithm to deal with this restriction.

REFERENCES

[1] Lowe, D., "Distinctive image features from scale-invariant keypoints, " *cascade filtering approach. IJCV 60*, pp. 91–110, 2004.

[2] Baumberg, A., "Reliable feature matching across widely separated views," *CVPR.*, pp. 774–781, 2000.

[3] Bay, H., Ess, A., Tuytelaars, T., and Gool, L., "SURF: Speeded Up Robust Features, " *Computer Vision and Image Understanding (CVIU)*, Vol. 110, No. 3, pp. 346–359, 2008

[4] Foo, J. and Sinha, R., "Pruning SIFT for scalable near-duplicate image matching, " *Proceedings of the eighteenth conference on Australasian database*, pp. 63–71, January 30-February 02, 2007, Ballarat, Victoria, Australia

[5] Lucchese, L. and Mitra, S., "Color Image Segmentation: A State-of-the-Art Survey, " *Proceedings of the Indian National Science Academy (INSA-A)*, New Delhi, India, vol. 67, no. 2, pp. 207-221, 2001.

[6] Beucher, S., "The watershed transformation applied to image segmentation, " *Scanning Microsc Suppl*, 6:299–314, 1992.

[7] Osberger. W., and Maeder, A., "Automatic Identification of Perceptually Important Regions in an Image, " *14th International Conference on Pattern Recognition*, 1:701–704, 1998.
Beucher, S., "The watershed transformation applied to image segmentation, " *Scanning Microsc Suppl*, 6:299–314, 1992.

[8] Ray, S. and Turi, R., "Determination of number of clusters in K-means clustering and application in colour image segmentation, " *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*, pp. 137–143.

[9] Bai, X., Latecki, L. J., and Liu, W, "Skeleton pruning by contour partitioning with discrete curve evolution," *IEEE Trans. Pattern Anal.*, Mach. Intell. 29, 3 (Mar. 2007), pp. 449–462.

[10] Ke, Y. and Sukthankar, R., "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors, " *CVPR.*, pp.91–110, 2004.