Feature description using local neighborhoods✩

Man Hee Leea, Minsu Chob, In Kyu Parkc,∗

aComputer Graphics Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon 305-700, Republic of Korea
bWILLOW Project-Team, Département d’Informatique de l’École Normale Supérieure, ENS/Inria/CNRS UMR 8548, Paris, France
cDepartment of Information and Communication Engineering, Inha University, Incheon 402-751, Republic of Korea

A R T I C L E   I N F O

Article history:
Received 13 September 2014
Available online 3 September 2015

Keywords:
Feature description
Local neighborhoods
Similarity function
Graph matching
MRF

A B S T R A C T

Feature description and matching is an essential part of many computer vision applications. Numerous feature description algorithms have been developed to achieve reliable performance in image matching, e.g. SIFT, SURF, ORB, and BRISK. However, their descriptors usually fail when the images have undergone large viewpoint changes or shape deformation. To remedy the problem, we propose a novel feature description and similarity measure based on local neighborhoods. The proposed descriptor and similarity is useful for a wide range of matching methods including nearest neighbor matching methods and popular graph matching algorithms. Experimental results show that the proposed method detects reliable matches for image matching, and performs robustly to viewpoint changes and shape deformation.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Many computer vision techniques such as structure from motion [1], visual SLAM (simultaneous localization and mapping) [17] use multiple images or video frames for their input. To estimate camera motion or generate 3D maps in those applications, they usually extract local features, and then match/track them over a sequence of multiple images. Consequently, the performance is highly affected by the accuracy of feature extraction and matching algorithms used. Furthermore, a computationally efficient feature extraction is required to process a large number of high resolution images. During the last decade, excellent methods such as SIFT (scale-invariant feature transform) [21] and SURF (speeded up robust features) [5] have been developed and widely used for feature detection and description. Although they show robust performance when images have undergone in-plane rotation, scale variation, and illumination changes, they fail when the images involve severe variations caused by out-of-plane rotation of cameras, large viewpoint changes, or non-rigid object deformation. Note that those variations are often observed in real-world scenes, and thus should be handled by further research. In this paper, we propose a novel feature description method based on local neighborhoods, which effectively handles significant viewpoint changes and deformation of non-rigid objects. In the proposed approach, initial keypoints are first detected by existing feature detectors such as SURF [5], and then local graphs are constructed for each feature using its neighboring features. For robust matching, similarity between two features is evaluated based on their local graphs. Unlike previous patch-based descriptors [5,21], the proposed method provides a robust similarity function by taking into account neighboring features altogether. It is significantly more efficient than previous high-order graph matching [12] and progressive matching [10] techniques. The preliminary version of this paper was published in [19].

The main contributions of this paper are summarized as follows.

• We introduce a novel feature description based on local neighborhoods, and propose a similarity function that is robust to viewpoint changes and shape deformation.
• We demonstrate effective graph-based feature matching using the proposed similarity function.
• The proposed method is compared with recent popular feature descriptors, and shows superior performance in matching accuracy and efficiency.

This paper is organized as follows. In Section 2, we introduce related work. Section 3 describes the proposed feature description in detail. In Section 4, we describe feature matching based on a MRF formulation. Experimental results are presented in Section 5. We conclude with final remarks in Section 6.

2. Related work

2.1. Feature detection and feature description

Numerous feature detectors have been proposed in the literature. Mikolajczyk and Schmid [23] proposed Harris–Affine feature detector that is invariant to scale and affine transformations. They detect
interest points with a Harris–Laplace operator, that is adapted to affine transformations based on the second moment matrix. Agrawal et al. [2] proposed a scale-invariant detector (i.e. CenSurE) that is computed over multiple scales on the original image. They utilized the simplified bi-level kernels to achieve fast computation using a slanted integral image. Rosten et al. [25] proposed a fast and high quality corner detector (i.e. FAST) based on a segment-test algorithm. They applied a machine learning approach to implement the real-time algorithm, and generalized the detector with a ternary decision tree. Lowe [21] proposed a scale-invariant feature detector based on a multiscale extrema of DoG (difference-of-Gaussians) function. Alcantarilla et al. [4] proposed KAZE feature detection algorithm in a nonlinear scale space using nonlinear diffusion filtering rather than using Gaussian filtering. Dragon et al. [11] proposed NF-Features which represent object regions in the non-textured area.

On the other hand, feature descriptors also have been developed along with feature detectors. Lowe [21] proposed SIFT feature descriptor based on a histogram of oriented gradients. Currently, it is still widely used as the state-of-the-art method with high repeatability and accuracy. However, it requires a high computation cost in matching as the dimensionality of the descriptor is relatively high. Mikolajczyk and Schmid [24] proposed GLOH descriptor using gradient location and orientation histogram by extending the SIFT descriptor, which improves the robustness and distinctiveness using a circular sampling pattern. Bay et al. [5] proposed an accelerated version of SIFT (i.e. SURF) using an integral image and an approximated integer Gaussian filter. They do not resize the original image for multi-scale but simply increase the size of the filter. If it is applied to an integral image, Hessian determinant is calculated in a constant time. It is more robust to image noise than SIFT because they also integrate the gradient information within a subpatch. Alahi et al. [3] proposed FREAK descriptor inspired by the retina. They utilized a series of DoG functions over a retinal sampling pattern to generate the binary descriptor. Calonder et al. [7] proposed a fast and accurate descriptor (i.e. BRIEF) for real-time matching using short binary strings. They directly computed binary strings from image patches using a Gaussian sampling pattern and evaluated the Hamming distance for fast matching. However, it is neither rotation- nor scale-invariant. Stanski and Hellwich [29] proposed Spider descriptor which is characterized by a constellation of surrounding features.

Rublee et al. [26] proposed a fast feature detection and description algorithm based on the oriented FAST detector and the rotated BRIEF (binary robust independent elementary features) descriptor. Since BRIEF is rotation-variant, they calculated the orientation of FAST features using an intensity centroid scheme. Leutenegger et al. [20] proposed a fast and efficient feature detector and descriptor. They used scale space keypoint detection based on FAST in scale pyramid, and used a circular sampling pattern to generate a binary descriptor by performing simple brightness comparison.

Although these algorithms are robust in challenging conditions such as in-plane rotation and scale variations, most of them fail when images have undergone considerable out-of-plane rotation or shape deformation.

3. Feature description using local neighborhoods

3.1. Local graphs in image matching

Most image matching algorithms employ a patch-based feature descriptor to find initial correspondences between images. Some of them use a simple appearance-based matching algorithm as an initial step, and then apply geometric constraints to find reliable correspondences. Graph-based representation is widely used for object detection and recognition [14,15], where an object is modeled as deformable constellations of parts. Graph-based matching techniques to handle different objects in the same category or the same object with large view changes correctly. However, many of those algorithms require time-consuming cost minimization techniques for graph matching.

Unlike these previous methods, we propose a feature representation that encodes neighboring features relations. The local neighborhood has a larger supporting region than the patch-based descriptors, and preserves its structure not only for large view changes but also for deformable objects. These feature description can be used for robust matching.

Fig. 1 compares a conventional patch-based descriptor with the proposed local neighborhood representation, where initial keypoints are detected by SURF. The images on the left in Fig. 1 show local patches at the different keypoints. The local patches are inconsistent to each other because the target objects are different, even though they are from objects in the same category. It means that
an appearance-based matching technique is unlikely to find correct correspondences based on the patch-based descriptors. On the other hand, the images on the right show the corresponding local graphs at the keypoints, which are constructed using the neighboring features. As salient features have high repeatability even in case of deformation or large view changes, local graphs constructed using such salient features are likely to preserve their graph structures as shown in this example. Each local graph is organized with a root and leaf node. Each leaf node in the proposed local graph encodes normalized distance and the related orientation from the root node.

3.2. Modeling local graphs

To make a robust feature descriptor, we construct a local graph for each feature as summarized in Algorithm 1. Keypoints $\mathcal{K} = \{v_i\}_{i=1}^M$ and $\mathcal{K}' = \{v'_i\}_{i=1}^M$ are detected from input images $I$ and $I'$, and used as nodes for the local graphs. Conventional algorithms for corner or blob detection can be used for feature detection. In our approach, FAST and SURF detectors are employed as they have good performance in efficiency and repeatability under affine transform. Using the detected features, a star-structured graph $G$ is constructed for each feature by exploiting the product of the prior probability $\psi_i(x_i)$ of the prior term encodes spatial smoothness.

The goal of feature matching is to find the best correspondences between two keypoint sets $\mathcal{K} = \{v_i\}_{i=1}^M$ and $\mathcal{K}' = \{v'_i\}_{i=1}^M$. In the proposed matching procedure, the correspondences are acquired using two graph sets $\Omega = \{G_i\}_{i=1}^M$ and $\Omega' = \{G'_i\}_{i=1}^M$. To find the optimal correspondences, traditional distance based feature matching methods such as distance thresholding or ratio test can be applied. Furthermore, feature matching can be formulated as a Bayesian framework using MRF. Two complete graphs $\Omega$ and $\Omega'$ are constructed using the root features of each local graph in $\Omega$ and $\Omega'$. The unknown variable $x$ is defined as a multi label vector $x \in \{0, 1, \ldots, M\}^M$. If local graph $i$ in $\Omega$ matches local graph $j$ in $\Omega'$, then $x_i = 1$ where $x_i$ is the ith element of vector $x$. The posterior probability $p(x | \Omega, \Omega')$ is proportional to the product of the prior probability $p(x)$ and the likelihood probability $p(\Omega, \Omega' | x)$. The objective function of $x$ is defined as posterior energy $E(x)$ consisting of the unary term and the pairwise term.

We use the similarity function of Eq. (1) for the unary energy function $\psi_i$:

$$\psi_i(x_i) = \exp\{-S(i, x_i)\}$$

(2)

The unary energy is defined by using the similarity between the local graphs $G_i$ and $G'_i$.

On the other hand, the pairwise energy function $\psi_{ij}(x_i, x_j)$ is simply defined as

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } v_i \in \mathcal{V}_i \text{ and } v'_j \in \mathcal{V}'_j \\ \alpha & \text{otherwise} \end{cases}$$

(3)

where $\alpha$ is a non-negative constant. If local graph $G_i$ of $v_i$ includes $v_i$ and local graph $G'_i$ of $v'_i$ includes $v'_i$, the pairwise energy $\psi_{ij}(x_i, x_j)$ will be 0. This pairwise energy term encourages two nodes of the pair on the same domain to be in the neighborhood of each other. It means that the prior term encodes spatial smoothness.

Finally, the energy function is defined as

$$E(x) = \sum_i \psi_i(x_i) + \sum_{ij} \psi_{ij}(x_i, x_j).$$

(4)
The solution is obtained by $\mathbf{x}^* = \arg\min_\mathbf{x} E(\mathbf{x})$ using the max-product loopy belief propagation algorithm [13].

5. Experimental results

The performance of the proposed descriptor is evaluated and compared with recent popular feature descriptors, including SIFT [21], SURF [5], ORB (oriented FAST and rotated BRIEF) [26], and BRISK (binary robust invariant scalable keypoints) [20]. The original implementations of these algorithms are provided by the authors directly or indirectly via their OpenCV contributions. The proposed algorithm is implemented in C++ code based on the OpenCV interface.

5.1. Feature matching under significant viewpoint change

To demonstrate robustness of the proposed approach under large viewpoint change, we take the test images of an indoor scene (a bookshelf filled with books) with out-of-plane rotation by every 5 degrees between -70 and 65 degrees as shown in Fig. 2. We detect approximately 1,000 features in each image and match each rotated image to the reference image without rotation. Fig. 3(a) shows the percentage of correct matches among total matches, which indicates that the proposed method has more inliers than the others for most of the rotation angles. When the number of inliers is compared in the experiment, the proposed descriptor can match inliers 16.9% more than SURF, and 9.7% more than SIFT in average. Fig. 3 (b) shows the matching score of each algorithm across different viewpoint angles using Mikolajczyk’s evaluation method. In this experiment, we used the nearest neighbor matching method to find correspondences. The result indicates that the proposed local graph is more robust to large viewpoint changes.

For further experiments, we use Mikolajczyk and Schmid’s dataset [24]. The dataset has a variety of different imaging conditions including in-plane and out-of-plane rotation, scale and illumination change, image blur, and JPEG compression. The number of correct matches is counted by reprojection each feature onto another image using a ground-truth homography matrix. To evaluate the performance, we show recall vs. 1-precision plots using the graffiti, bricks, and boat images that have both in-plane and out-of-plane rotation. We detect approximately 1,000 features in each image and count the number of inliers and outliers. Fig. 4 shows comparative examples of the proposed method and the SIFT method under the same condition. The proposed approach finds significantly more inliers and less outliers than the SIFT method.

Fig. 5 shows the recall vs. 1-precision comparison of Fig. 4. The proposed approach significantly outperforms the others for the scenes with out-of-plane rotation and large view changes (graffiti and bricks). Furthermore, the proposed approach still shows better performance for the scene with in-plane rotation (boat). Table 1 presents computation time of each descriptor, which shows that proposed descriptor is faster than SIFT descriptor and comparable to SURF descriptor. In the matching procedure, the proposed method is much slower than binary descriptors but still comparable to the patch based descriptors. Note that, each node in the proposed descriptor includes the relative orientation, normalized distance, and feature index. The proposed method spends only 39% of SIFT memory consumption when 16 local neighbors are used.

5.2. Feature matching for non-planar scene and deformable object

The proposed method is tested on non-planar scenes, which include several 3D objects with ±45 degree viewpoint changes. As shown in Fig. 6, the proposed approach finds many more correct correspondences than the other algorithms.

We further test the proposed approach on the multi-view datasets [30]. Although they provide the camera projection matrix of each image, it is not suitable to determine the ground truth matching with 2D homography transformation as the objects in the dataset are not planar. Therefore, we compute the depth map of the source image using the ground truth 3D triangle model, and calculate the
Table 1  
Computation time with 1000 feature points in milliseconds. Intel Core i7 2.7 GHz processor with 16 GB memory is used.  

<table>
<thead>
<tr>
<th>Descriptor generation</th>
<th>Nearest neighbor matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>406</td>
</tr>
<tr>
<td>SIFT</td>
<td>1866</td>
</tr>
<tr>
<td>SURF</td>
<td>364</td>
</tr>
<tr>
<td>ORB</td>
<td>42</td>
</tr>
<tr>
<td>BRISK</td>
<td>51</td>
</tr>
</tbody>
</table>

Fig. 5. Recall vs. 1-precision plots for each algorithm. Horizontal and vertical axes denote 1-precision and recall, respectively. (a)–(c) Graffiti dataset. (d)–(f) Bricks dataset. (g)–(i) Boat dataset.

Figs. 8 and 9 show results of image matching with deformable surfaces on the EPFL dataset, in which the proposed descriptor is compared with KAZE [4], SIFT, and SURF. We detect 600 features from each image and find 400 correspondences using a distance

3D positions of the features. If distance of the matched feature in the target image to the projected point is less than a threshold, it is regarded as an inlier match. Fig. 7 shows the matching score of each algorithm for fountain-P11 and Herz-Jesu-P8 dataset. For the fountain-P11 dataset, the proposed descriptor matches inliers 22.8% more than SURF, and 15.0% more than SIFT in average.

Fig. 6. Visual comparison of matching results between the proposed algorithm and SIFT for toys dataset. Green lines show correct correspondence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).
thresholding. Even on these image pairs with considerable deformation, the proposed approach outperforms the others.

5.3. Graph based image matching

To evaluate how the proposed descriptor improves the actual image matching, it is tested on the graph based image matching scheme [10] for the VOC 2010 dataset. Since the proposed descriptor aims at robustness for deformable objects, different objects in the same category dataset can simulate these conditions. In this experiment, MSER [22] is used for the feature detection and we apply the proposed descriptor for the initial matching. The initial active graphs for graph matching is constructed from the initial candidates. Fig. 10 shows conventional one-shot graph matching results using the initial active graphs. More inliers are found than the original algorithm when the proposed descriptor is used for finding initial candidates. Fig. 11 shows that the proposed approach has better performance not only for one-shot graph matching but also for progressive graph matching.

6. Conclusions

We proposed a robust feature description method based on local neighborhoods that handles large viewpoint change and shape deformation. In the proposed approach, each feature consists in a local graph with neighboring features, and the local graph is used for a robust similarity function in image matching. Experimental results demonstrate that the proposed approach has better performance than the existing methods such as SIFT, SURF, ORB, and BRISK.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2013R1A2A2A01069181).
References


