Robust Invariant Feature Extraction for Object Recognition and Natural Landmark based Autonomous Navigation

Zhe Lin, Sungho Kim and In So Kweon
Robotics & Computer Vision Laboratory
Dept. of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology (KAIST)
373-1, Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
{limcher,shkim}@rcv.kaist.ac.kr, iskweon@kaist.ac.kr

Abstract

In this paper, we present a new robust image feature detector for the object recognition and vision based mobile robot navigation. The proposed algorithm extracts highly robust and repeatable features based on the key idea of tracking and grouping multi-scale interest points and selecting a unique representative structure with the strongest response in both spatial and scale domains. Weighted Zernike moments are used as the local descriptor for feature representation. The experimental results and performance evaluation show that our feature detector has high repeatability and invariance to large scale, viewpoint and illumination changes. The efficiency and usefulness of the proposed feature detection method are also confirmed by the excellent performance on object recognition and robot navigation.

1. Introduction

Recently, the use of local features in the context of recognition has been successful due to their invariance and power to handle occlusions and background clutters. Local features, physically, are characteristic local image structures with some meaningful information. Local feature extraction process is composed of interest point detection, invariant neighborhood extraction and local description. Figure 1 illustrates an example of the process of local feature extraction and matching.

The breakthrough work [1] on local features addresses the use of local invariants for content based image retrieval, and this has proved to be a very influential approach. Later, many researches have been done based on this framework or its generalizations. Lindeberg has proposed an automatic scale selection mechanism which finds 3D maxima of Laplacian of Gaussian filter in the scale space representation [2]. Similarly, Lowe has presented a computationally efficient feature detector – Scale Invariant Feature Transform (SIFT) [3] by searching for the 3D maxima of Difference of Gaussian filter in the pyramid scale space. Original SIFT has been further refined to reduce the noise sensitivity and edge effect in [4]. Following these works, Mikolajczyk and Schmid have proposed a Harris corner based scale invariant interest point detector [5]. The detector searches for the maximum response of Laplacian to estimate the characteristic scale. They later generalize this algorithm to the affine invariance by an iterative method [6]. Different from earlier works [7, 8], this method simultaneously adapts the location and scale instead of using fixed locations. The main drawback of these iterative approaches is high computational complexity. Slightly different techniques have been proposed in [9, 10] by directly extracting invariant regions based on local intensity information. But, for arbitrary scenes, the regions detected by these approaches are very inconsistent due to the large image intensity variations and only few of them can be matched.

Figure 1. An example of local feature extraction and matching process.
In this paper, we focus on the problem of extracting invariant regions and propose a new robust feature detector – Robust Invariant Feature (RIF) detector. It tracks interest points in the scale space to obtain group-wise feature representation of the image. And then, according to the information from each group of points, shape adapted invariant regions are extracted. Finally, rotation invariant weighted Zernike moments [11] are calculated on the normalized local image patches for the local description.

For application, RIF features are applied to object recognition and natural landmark based mobile robot navigation framework.

In the next section, we describe in the details the RIF region detector which is the main contribution of this paper. RIF description algorithm is given in Section 3. Section 4 and Section 5 introduce the object recognition and mobile robot navigation applications. Section 6 presents the experimental results for the proposed feature detector and its applications to recognition and robot navigation. Section 7 mentions about some of our work on generalization of RIF to affine invariance. Finally in Section 8, we summarize the paper with conclusion and future work.

2. RIF region detector

This section describes the proposed RIF region detector. Given images of various types of variations, the goal is to extract highly robust and repeatable regions so that they can be reliably applied to the image matching and recognition.

The algorithm first localizes interest points in the multi-scale image representation which is constructed with fixed incremental scale samplings. Next, it clusters these localized interest points to obtain the group-wise representation of interest points. Then, scale invariant characteristic local regions are determined by selecting scales at which the Harris measures reach the global maxima.

We are currently working on generalizing the developed algorithm to affine invariance so that the detector can handle larger viewpoint variations.

The main contribution of our work is that the proposed algorithm largely solves the location and scale ambiguity and noise sensitivity of the previous detector. Nearest neighbor searching, uncertainty region propagation and subsequent group-wise image representation assure the robustness and consistency of the detected local features.

2.1. Multi-scale interest points detection

Given an input image, we incrementally smooth the image to get the multi-scale representation. Since the intensive use of Gaussian filtering takes a lot of time, here, we uses the 4th order Recursive implementation of Gaussian filtering [12] which accelerates the large kernel filtering significantly.

After constructing multi-scale representation, the second moment matrix is calculated at every pixel location in the scale space images. Then, the normalized Harris measure is used to select interest points by finding its local maxima. For calculating the second moment matrix there are two scale parameters, the derivation scale \( \sigma_d \) and integration scale \( \sigma_i \).

Usually a constant factor \( s \) is used between the two scales to balance the scope of the area used to calculate the statistics of local gradient variation. The second moment matrix \( M \), Harris measure \( R \) and the two scales \( \sigma_d \), \( \sigma_i \) used in the proposed algorithm are as follows:

\[
M = \sigma_d^{-2} G(\sigma_i) * \begin{bmatrix}
L_i^2(x, \sigma_d) & L_iL_y(x, \sigma_d) \\
L_iL_x(x, \sigma_d) & L_y^2(x, \sigma_d)
\end{bmatrix}
\]

\[
R = \det(M) - \alpha \text{Trace}^2(M) \quad \sigma_i = \sigma_0 \sigma^n \quad \sigma_d = s \sigma_i
\]

In (1), the parameters are selected as initial scale \( \sigma_0 = 1.0 \), scale factor \( \sigma = \sqrt{2} \), the number of scales \( N = 10 \) and constant factors \( \alpha = 0.06 \), \( s = 0.7 \). Figure 2 shows an example of the multi-scale interest point detection. Interest points detected from different scales are marked by different color pixels. We can find out some important facts from the above results:

a. Number of interest points decreases when scale increases.

b. Interest point locations vary slightly while scale increase, the larger the scales, the far the distance of interest points from the true corners.

c. Locus of the interest points is relatively constant to the rotation, scale, viewpoint and illumination changes.

Figure 2. Multi-scale interest points.
We can see from figure 2 that interest points can be group-wise clustered so that each group represents a local structure. The advantage of this representation is that the trace of the interest points corresponding to each structure is relatively stable under image variations. We define this scale space trace of the interest points as \textit{corner evolution}. With the scale increasing, interest points from different evolution group can largely intersecting. It gives rise to a question that how to separate these points or how to assign them to each local structure. For solving this ambiguity, we propose following tracking algorithm which correctly classifies points to each structure.

2.2. Tracking and grouping

The algorithm is as follows:

i) For initial scale image, assign a group for each interest point.

ii) From the next scale level \( L \), for each interest point \( P \) in this level, search for the corresponding point link in the lower level \( (L-1) \).

\text{IF} any corresponding link exists,  
\text{THEN} assign \( P \) to this link and group.  
\text{ELSE} assign this interest point to a new group.

iii) Go to next scale level \( L+1 \), and iterate ii \sim iii.

The idea is to cluster those multi-scale interest points corresponding to the same local structure. The linking is propagating from the highest scale level to the lower scales based on the principle of nearest neighbor searching within the uncertainty regions. The tracking of interest points continues until no corresponding link points within the uncertainty region. Here the uncertainty region propagation is very important as it highly affects the grouping results. For example, when many textures existing in the image, the relative distances between local structures can be very small, so, many false grouping may be generated, which consequently decreases the number of correct estimations. Because interest point propagation in the scale space is directly related on the scale propagation, we model the corner propagation as the following:

\[ R(s, L) = k \cdot s^s \]  
(2)

Where, \( R(s, L) \) represents the radius of uncertainty region, \( k \) is a constant factor, \( s \) is the parameter of the function and \( L \) is the scale level index. The parameter \( s \) is chosen as the same as the scale factor \( \sigma \). For the parameter \( k \), we have tested various values with respect to the resulting grouping error (See table 1). The best value is obtained as \( k = 0.5 \). Figure 3 shows the grouping results when \( k = 0.5 \), the result is very consistent to image transformations such as scale, viewpoint and illumination changes.

<table>
<thead>
<tr>
<th>( k )</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grouping Error (%)</td>
<td>10.4</td>
<td>4.8</td>
<td>1.2</td>
<td>2.5</td>
<td>3.9</td>
<td>5.2</td>
<td>7.3</td>
</tr>
</tbody>
</table>

Figure 3. Example: grouping results under scale and viewpoint changes.
2.3. Shape adaptation

Our algorithm detects a unique characteristic scale in each group by using the group-wise feature representation. Consequently, the ambiguity and inaccuracy in detected features can be reduced maximally. Our scheme is to select a unique representative scale at the strongest peak point so that the corresponding region appears the most corner-like structure. This is similar to the visual perception on corner structures since many evidences from neuropsychology and cognitive science have shown that human visual system has stronger attention or response in the high curvature points or corners than edges and other features. We can further estimate the exact scale by fitting parabola to each selected peak (3).

\[ S' = \frac{S_{n+1} + S_{n-1}}{2} - \frac{\sigma(S_{n+1} - S_{n-1})R_{n+1}^2 - R_{n-1}^2}{R_{n+1}^2 - (\sigma + 1)R_{n-1}^2 + \sigma R_{n+1}^2} \] (3)

Then, the feature point is re-localized based on the estimated scale \( S' \) and the sub-pixel accuracy in localization is obtained by weighted average of contributions from eight neighboring points.

3. RIF descriptor

For description, we first normalize the extracted patches to the canonical 1010 × 10 circles. Then, dominant orientation is estimated using gradient distribution on the local image patch [6]. We assume the conventional linear illumination change model as follows:

\[ I' = sI + o \] (4)

Based on this model, we normalize the image patches by linearly changing their mean and variance to a fixed value. In this way, the scale \( s \) and offset \( o \) can be eliminated to get illumination normalized patch. In our feature, weighted Zernike moments [11] are used. Zernike moments have superior properties in terms of image content representation, information redundancy and noise characteristics [14] so they can be reliably used in recognition problem. It is defined over a set of complex polynomials which form a complete orthogonal set over the unit disk. Zernike moments are calculated by projecting the image intensity onto these orthogonal basis functions. Gaussian window is used to weight the image patch before calculating the descriptors. Euclidean distances are used as the similarity measure for matching. The weighted Zernike moments are calculated as follows:

\[ A_{mn} = \frac{R + 1}{\pi} \sum_{\alpha=1}^{R+1} \sum_{\beta=0}^{R+1} \omega(W(x,y)f(x,y)|V_{mn}(x,y)|^2) \] (5)

\[ V_{mn}(x,y) = R_{mn}(x,y)\exp(jm\tan^{-1}(y/x)) \]

\[ R_{mn}(x,y) = \sum_{s=0}^{(n-m)^2} (-1)^s \binom{n-s}{m} \frac{(n-s)!}{2^{n-s}} \frac{(-1)^{s/2}}{s!} \frac{(x^2 + y^2)^{s/2}}{2} \]

Where, \( V_{mn}(x,y) \) is orthogonal basis function, \( R_{mn}(x,y) \) is radial polynomial, \( f(x,y) \) is image function and \( W(x,y) \) is the Gaussian weight. We use the fast implementation [15] to calculate Zernike moments.

4. Object recognition and pose estimation

The framework of the object recognition system is drawn in figure 4. It consists of on-line and off-line parts. In the off-line learning stage, RIF features are detected from the model images and the resulting descriptors are stored in the database. In the on-line recognition stage, the features from the input and model images are pushed into the searching engine to find the most similar model. The Euclidean distance is used to measure the similarity between descriptors. We use Approximate Nearest Neighbors (ANN) search algorithm and probabilistic voting technique for efficient DB indexing. Final verification stage ensures the recognition result to be correct by estimating the optimal homography and counting inliers and outliers.

![Figure 4. Object recognition system framework.](image)

5. Robot navigation

The key to the navigation is the estimated pose information. We first manually drive the robot through the environment and capture images at critical locations that the robot is expected to do turning motion. And then, the group of RIF features are
extracted from these DB images as the scene landmarks. Finally through our recognition and pose estimation system, the robot motion is planned to iteratively correct its motion and converges to the optimal pose matching with database pose. We have used KASIRI-IV robot as the experimental platform which is wireless commanded network based robot driven by Strong Arm embedded system. The iterative pose converging is based on the estimated landmark ID and homography and verified optimal affine transformation. The homography and estimated optimal affine transform is as follows:

\[
H = \begin{pmatrix}
 h_{10} & h_{11} & h_{12} \\
 h_{20} & h_{21} & h_{22} \\
 0 & 0 & 1
\end{pmatrix} \rightarrow \left( \begin{array}{c} A' \\ t \end{array} \right) \quad (6)
\]

The motion command is given by translation vector and relative scales through iteratively converging \( A' \) to the identity matrix.

6. Experimental results

In this section, we present experimental results for the proposed feature detector and its application to object recognition and mobile robot navigation.

6.1. Qualitative evaluation

We have tested the performance of the proposed detector under various image variations. The Hangul image set is selected as the experimental samples. The results show that the RIF features are very consistent and robust to scale and illumination changes (figure 5). For viewpoint variation, features are consistent in a certain range of viewing angles (\( \theta < 40' \)).

6.2. Quantitative evaluation

We use the repeatability criterion [13] to evaluate the performance of the RIF region detector. The sequences used in the experiment are selected from INRIA image database. Figure 6(a) shows that our proposed detector obtained higher repeatability than SIFT and Harris-Laplace (H-L) detector over the scale range of 1 to 4. The evaluation under viewpoint change shows that our detector has a slightly better repeatability rates than Harris-Laplace detector (figure 6(b)). Although our detector shows a high repeatability and robustness, it is still very sensitive to the large viewing angle changes. Our current work is focusing on the affine generalization in order to cover large viewpoint changes.

![Figure 6](image-url-a.png)  ![Figure 6](image-url-b.png)

Figure 6. Repeatability evaluation of the proposed detectors and SIFT & H-L detector. (a) To scale changes. (b) To viewpoint changes.

![Figure 7](image-url-a.png)  ![Figure 7](image-url-b.png)

Figure 7. Recognition results under large scale and viewpoint changes.

<table>
<thead>
<tr>
<th>Recognition Rate (%)</th>
<th>Scale changes</th>
<th>Viewpoint change</th>
<th>Illumination changes</th>
<th>Occlusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIF</td>
<td>63/63 (100%)</td>
<td>115/120 (95.8%)</td>
<td>80/80 (100%)</td>
<td>60/60 (100%)</td>
</tr>
<tr>
<td>Multi-scale Harris Detector</td>
<td>61/63 (96.8%)</td>
<td>116/120 (96.7%)</td>
<td>80/80 (100%)</td>
<td>60/60 (100%)</td>
</tr>
<tr>
<td>SIFT</td>
<td>63/63 (100%)</td>
<td>100/120 (83.3%)</td>
<td>76/80 (95%)</td>
<td>56/60 (93.3%)</td>
</tr>
</tbody>
</table>
6.3. Recognition performance

Finally, the performance of the proposed feature detector is evaluated by applying it to the recognition system. We tested the recognition performance for the 20 different classes of total of 483 object images. The proposed detector shows very robust results to the rotation, scaling, viewpoint variation, illumination change, occlusions and background clutters.

We have tested the recognition performance on the 20 different classes of 483 object images which are chosen from the KAIST DB. Figure 7 shows that proposed feature detector can handle large scale and large viewpoint changes. Quantitative performance is evaluated for all the database images and the percentage of correct recognition (Recognition Rate (%)) is compared with other two feature based approaches (table 2). As a result, our features show the best recognition performance under scale changes, illumination changes and occlusions. For the view change sequence, our detector shows slightly lower recognition rate 95.8%.

6.4. Navigation performance

Figure 8 shows example results of object recognition, pose estimation from a indoor navigating robot. We can see the system correctly recognizes scene landmark even in a severe occlusion and scale changes.

In our navigation experiment, the robot correctly achieved its goal to the final destination with 90% success rate among over 20 trials. Currently, we are generalizing this system to an automatic map learning and localization using the additional wheel odometry information.

7. Generalization to affine invariance

We have analyzed the evolution properties of eigenvectors of the second moment matrices in the scale space. Figure 9(a) shows the separated scale selection by eigenvalues and figure 9(b) shows the selected scales plotted on the images. We consider each local structure as a twin-Gaussian field where optimal elliptical kernel can be estimated by searching for the highest local energy between regions induced by the two estimated scales. From this assumption, we can estimate the optimally interpolated elliptical kernels and updated feature locations as follows:

\[
\sigma_{\text{shift}}^* = \sigma^* + b/2 \quad lx = \text{ratio} \cdot px + (1 - \text{ratio}) \cdot qx
\]
\[
\sigma_{\theta}^* = \sigma^* + b \cos \theta / 2 \quad ly = \text{ratio} \cdot py + (1 - \text{ratio}) \cdot qy
\]
\[
\text{ratio} = \sigma_1 / (\sigma_1 + \sigma_2)
\]

Note that the above result (figure 9(c)) is very preliminary. We are continuing to work on how to optimally adapt the local shape and obtain affine invariance.

Figure 8. Navigation examples. (a) Landmarks. (b) Recognition under occlusion. (c) Pose estimation under scale and view changes.

Figure 9. Anisotropic shape adaptation.
8. Conclusion and future work

We have developed a new robust local feature detector and showed its application to object recognition and natural landmark based mobile robot autonomous navigation. Our detector, from the scale space interest point propagation to tracking and scale adaptation, basically models the fundamental knowledge and rules on visual corner perception. Various experimental results have shown that our detector can generate highly robust and reliable features which can be efficiently used in the recognition. Although our proposed feature detector showed a good performance, more precise scheme of tracking and peak modeling is desirable so that the performance can be further improved. As the proposed RIF detector can not handle large viewpoint changes due to the large distortion of scale invariant circular regions, we are currently generalizing the algorithm to the affine invariant case which can handle large viewpoint changes. Another further work is to develop an automatic map learning algorithm and eventually to build a fully autonomous indoor SLAM system.

Acknowledgement

This research has been partly supported by Ministry of Information and Communication, Korea Industrial Technology Foundation and National Research Laboratory (NRL: M1-0302-00-0064) of Ministry of Science and Technology.

References