Evaluation of Local Features for Near-Uniform Scene Images

Tze Kian Jong
Faculty of Engineering
Universiti Malaysia Sarawak
Sarawak, Malaysia
johntkian@gmail.com

David B. L. Bong
Faculty of Engineering
Universiti Malaysia Sarawak
Sarawak, Malaysia
bbldavid@unimas.my

Abstract— Image stitching requires accurate matching of visual features to achieve good alignment. However, feature-based matching often has poor result particularly when image content is fairly near-uniform and thus it remains a challenging problem to be addressed. When the current state-of-the-art feature detectors unable to detect sufficient reliable corresponding keypoints, the output stitched images often suffer from misalignment, projective distortion and visible artefact. This paper presents a new experimental evaluation using especially near-uniform images for the performance of some well-known feature detectors, such as Harris, SIFT, SURF, BRISK and KAZE. In addition, we have also introduced RCS-Score to compare spatial distribution of the correct matched keypoints in overlapping region between images. The results show that the best performed local feature detector is KAZE. However, none of the tested feature detectors can reach more than 50% spread of the overlapping region.

Keywords— local features, interest points, image stitching, near-uniform scenes, quantitative evaluation

I. INTRODUCTION

Image stitching is basically a process of merging images with overlapping fields of view in order to obtain an image with wider field of view and higher resolution. It is fundamentally well studied in computer vision [1]. The attempt to produce image stitching with satisfactory visual results typically involve two main approaches: (1) image alignment [2, 3, 4, 5], and (2) image composition such as seam cutting [6] and blending [7]. Image alignment is considering the most crucial step in image stitching. Its execution depends very much on the task of finding precise corresponding keypoints between images of the same overlapping scene or object. When the current state-of-the-art feature detection algorithm incapable to detect sufficient and reliable correspondences, the remaining steps of image stitching will fail to deliver a pleasing stitched image and resulting in artefacts, for example misalignment or ghosting visual. It is especially obvious when image content is nearly uniform or low-texture that hardly to extract satisfactory number of correspondences in order to bring the overlapping scene into exact alignment. This is due to the homogeneous regions in near-uniform or low-texture scene images, such as sky, ocean, coast, land surface etc., that are not distinctive enough to provide rich and reliable correspondences.

In addition, irregular spatial distribution of the corresponding keypoints within overlapping region also triggering misalignment, distortion and visible artefacts in the stitching result. For instance, Fig. 1 (a) and Fig. 1 (b) depict the correct matched SIFT and KAZE features respectively between temple image pair (images obtained from [2]). Notice that both set of matched correspondences are detected in different area of the overlapping region. To fit each set of data, this will lead to estimation of two Unlike global homography model and consequently producing misalignment of the final stitched images. Fig. 1 (c) and Fig. 1 (d) show obvious misalignments (see red rectangles) in the stitching results. This is understandable that estimation of the global

![Matched SIFT features](image1)

![Matched KAZE features](image2)

![Image stitching using matched SIFT features](image3)

![Image stitching using matched KAZE features](image4)
homography model is computed based on only the data that appear within a certain area of the overlapping region (compare Fig. 1 (a) and Fig. 1 (b)). The effects of the estimated homography model will further degrade the stitching result when reliable corresponding keypoints are insufficient within the overlapping region.

It is thus of our great interest to investigate how well is the current advance feature detectors perform in stitching near-uniform-scene images. In this paper, we first detect and match the Harris, SIFT, SURF-64, SURF-128, BRISK, KAZE-64 and KAZE-128 features in the near-uniform scene images and then stitch them together. SURF and KAZE feature descriptors are differed by 64 and 128 variants to specify its dimension of description vectors, with larger feature size provides greater accuracy in matching process. Then, we study the spatial distribution uniformity of corresponding keypoints within overlapping region using an improved version of spread measure. Finally, we suggest the most appropriate feature detectors for matching and stitching near uniform scene images. This paper is organized as follows. Section 2 reviews the related work of some advance feature detectors and evaluation methods. Section 3 describes image database and algorithm methods used in this experiment. Results and discussions are presented in Section 4. Finally, we summarize the conclusions in Section 5.

II. RELATED WORK

A. Feature Detectors and Descriptors

The challenges in image stitching is obviously in its ability to accurately detect image features. Many well-known feature detectors and descriptors algorithm were proposed by previous researchers. They can be tracked back to the work of Harris corner detector in 1988 to detect both corner and edge features based on the second-moment of eigenvalues matrix [8]. But Harris corners are not scale invariant. The most popular multiscale feature detection and description algorithms was proposed by Lowe in 2004 to describe the Scale Invariant Feature Transform (SIFT) keypoints [9, 10]. SIFT has proven remarkably effective and it is widely used in numerous visual feature-based applications, including object recognition, image matching and stitching, visual mapping etc. Though, it is not suitable for real-time systems as it imposes a great computational demand. This has led to an intensive research for replacements with lesser computation complexity while upheld the methods performance as in consistent with SIFT. In 2006, Bay et al. proposed the Speeded up Robust Feature (SURF) that has similar matching rates to SIFT with much faster performance [11]. Later in 2011, Leutenegger et al. proposed Binary Robust Invariant Scalable Keypoints (BRISK) that shows comparable performance and much faster to compute than SIFT and SURF [12]. BRISK features are scale and rotation invariant. In the context of scale space, both SIFT and SURF approaches make use of the Gaussian scale space that do not preserve object boundaries. To overcome this problem, Alcantarilla et al. introduced KAZE features in 2012 to detect and describe features in nonlinear scale spaces which makes blurring locally adaptive to small image details but preserving object boundaries [13].

B. Quantitative Evaluation

Evaluation performance of local features is based on the number of correct matches and their spatial distribution within overlapping region obtained for an image pair. This section describes the evaluation method used in this study.

In order to match two planar scene images, a detected keypoint \( x_i \) in image \( I_i \) should have a corresponding keypoint \( x_j \) repeated in image \( I_j \). According to [14, 15, 16], repeatability \( R(e) \) is the ratio of number of keypoints repeated between images \( I_i \) and \( I_j \) with respect to the minimum number of detected keypoints, as defined in equation (1).

\[
R_i(e) = \frac{|C_i(e)|}{\min(n_i, n_j)}
\]

where \( C_i \) denotes the number of corresponding keypoints, \( n_i \) and \( n_j \) are the number detected keypoints in images \( I_i \) and \( I_j \) respectively. A repeated keypoint \( x_i \) is not exactly detected at position \( x_j \), but rather in \( x_j \) neighbourhood of size \( e \). Hence, \( C_i(e) \) is determined if the error in relative location does not exceed \( e \) size neighbourhood of \( x_j \) [16]. In our evaluation, instead of \( e \), we use a threshold \( t < 1.5 \) distance between descriptors to determine correspondences.

As defined in equation (2), recall \( RC \) is the number of correct matches with respect to the number of correspondences between two images of the same scene [17]:

\[
RC = \frac{|M_j(e_s)|}{|C_j(e)|}
\]

where \( C_j \) denotes the number of corresponding keypoints, \( M_j \) is the number correct matched keypoints. In [17], the number of correct matches is determined by the overlap error \( e_s < 0.5 \) in image area that covered by two corresponding regions. As an alternative, we use M-estimator SAmple Consensus (MSAC) algorithm [18] to exclude outliers and determine the correct matches based on distance error \( e_d \) (< 1.5 pixel) from a keypoint to its projection corresponding keypoint between images.

The spread measure was inspired by the similar manner done by Andres Marmol et al. to compute spatial distribution of the detected keypoints in an image [14]. Instead of using a uniform 10 × 10 grid cell recommended by Andres Marmol et al., we divided an image into square grid cells with area of each square is 0.25% of the image total area. The number of valid grid cells containing at least one keypoint in relation to the total number of grid cells of an image is represented by the spread \( S \):

\[
S = \frac{n_g}{N_g}
\]

where \( n_g \) is the number of valid grid cells contain at least one keypoint, and \( N_g \) is the total number of square grid cells of an image. Besides, to quantify the spread-overlap \( S \), for the overlapping region between two images, we use the same equation (3) by computing the number of valid grid cells containing at least one keypoint with respect to the total number of square grid cells covering the overlapping region.

III. EXPERIMENTAL METHOD

A. Image Datasets

Performance of the current state-of-the-art feature detectors are evaluated on some selected standard datasets (obtained from [2, 17, 19, 20]) and real images of different scene types that partially illustrate near-uniform or low-texture content. Fig. 2 shows the example of datasets used in our experimental evaluation: (1) two set of planar scenes, each comprises six sequences of blurred images; (2) four standard datasets of image pairs; and (3) two set of real image pairs.
A. Performance evaluation method.
B. blurring amount in relation to the behaviour of correct matches

image blur datasets are used initially to study the influence of image content become smoother and homogenous. Thus, the blurry, feature

categorized as near

captured by authors. Notice that Fig. 2(a) and Fig. 2(b) are not categorized as near-uniform scene but they are sequence of blurred images. When an image undergoing increasingly blurry, feature detection is less accurate or sensitive because image content become smoother and homogenous. Thus, the image blur datasets are used initially to study the influence of blurring amount in relation to the behaviour of correct matches in overlapping region of images.

B. Algorithms

This section reveals the algorithms used to evaluate local features in our study. The first algorithm describes image stitching method, whereas the second algorithm explains the performance evaluation method.

Algorithm 1: Image stitching

<table>
<thead>
<tr>
<th>Input: Images $I_i$, $i = 1, \ldots, N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Compute local feature keypoints</td>
</tr>
<tr>
<td>for $i = 1$ to $N$ do</td>
</tr>
<tr>
<td>a. Detect Harris, SIFT, SURF-64, SURF-128, KAZE-64 or KAZE-128 keypoints</td>
</tr>
<tr>
<td>b. Compute keypoints description</td>
</tr>
<tr>
<td>end for</td>
</tr>
<tr>
<td>2. Match keypoints using vl_ubcmatch.m [21]. Default threshold, $t = 1.5$. Return indexed matches</td>
</tr>
<tr>
<td>3. Remove outliers using MSAC algorithm. Number of random trials set to 500. $\epsilon_d = 1.5$.</td>
</tr>
<tr>
<td>4. Estimate geometric transformation (projective type) based on corresponding inliers</td>
</tr>
<tr>
<td>5. Apply geometric transformation to warp images</td>
</tr>
<tr>
<td>6. Blend and overlay warped images</td>
</tr>
<tr>
<td>7. Compute performance evaluation (see Algorithm 2)</td>
</tr>
<tr>
<td>Output: Stitched image pair, detected keypoints, corresponding keypoints and correct matched keypoints</td>
</tr>
</tbody>
</table>

Algorithm 2: Performance evaluation

<table>
<thead>
<tr>
<th>Input: Images $I_i$, $i = 1, \ldots, N$ and set of detected, corresponding and correct matched keypoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Compute repeatability $R$ and recall $RC$ using equation (1) and (2)</td>
</tr>
<tr>
<td>2. To compute spread $S$ over an image, define square grid dimension (row $r = column c$) based on 0.25% of image $I_i$ area</td>
</tr>
<tr>
<td>3. To compute spread-overlap $S_o$ over an overlapping region, define square grid dimension (row $r = column c$) based on 0.25% of the total overlapping area between two images $I_j$</td>
</tr>
<tr>
<td>4. Search and count $n_g$ and $N_\theta$</td>
</tr>
<tr>
<td>5. Compute $S$ and $S_o$ using equation (3)</td>
</tr>
<tr>
<td>Output: $R$, $RC$, $S$ and $S_o$</td>
</tr>
</tbody>
</table>

C. Computation

The experiment was conducted on a Windows 10 64-bit computer equipped with Intel core i5-6300U CPU @ 2.40GHz and 8.00GB of RAM. We utilize the MATLAB computer vision system toolbox in our study. The experiment also makes use of the VIFeat version 0.9.21 open source library for the SIFT features detection and description [21].

IV. RESULTS AND DISCUSSION

The aim of this study is to look at the performance of local feature detectors in near-uniform scene images. Fig. 3 shows the results for the sequence of planar scene images (from Fig. 2(a) and Fig. 2(b)) with each undergoing an increasing degree of blurring. Fig. 3(a) and Fig. 3(b) show the repeatability and recall measure for Fig. 2(a). When images become blurry, features repeatability and recall rate for all detectors are declining, except for the repeatability of BRISK features. BRISK suffers from detecting enough keypoints when image degradation (image blur) increase. Therefore, repeatability curve for BRISK is increasing according to equation (1) when number of corresponding keypoints are just slightly lower than the number of detected keypoints.

Fig. 3(c) shows how well do the correct matched keypoints spread over the overlapping region between images from Fig. 2(a). Higher percentage of spread-overlap $S_o$ means the correctly matched keypoints are well spread over the overlapping region that in turns benefit in reducing misalign and distortion in stitching result. The best results are obtained by the KAZE-64 detector with the highest 79.6% spread and lowest 31.7% spread in overlapping region when images become blurry. Fig. 3(d) shows another interesting curve that represents the division of recall with respect to spread-overlap (we call it $RC/S_o$ score) for overlapping region, in which we believe it would be useful as an alternate evaluation criterion to examine the performance of local feature detectors. In general, we noticed that $RC/S_o$ scores close to value one reflects better stitching result. This can be confirmed with the high $RC/S_o$ scores for Harris detector when it indeed fails to stitch Images 1–4 to 1–6 accurately (see Fig. 3(d)).

The effects of image blur for Images sequence 2 from Fig. 2(b) are presented in Fig. 4. Again, we can clearly see that repeatability curve for the BRISK detector is increasing due to its lowest keypoints detection rate as compare to other detectors when images become blurry. All detectors show
Fig. 3. Performance of local features over six sequentially blurred images (Images sequence 1 from Fig. 2(a)).

Fig. 4. Performance of local features over six sequentially blurred images (Images sequence 2 from Fig. 2(b)).

decreasing recall curve in Fig. 4(b) indicating features detection accuracy are in fact affected by images blur. The results from Fig. 4(c) and Fig. 4(d) also show that both Harris and BRISK detectors perform worst with less than 20% spread-overlap and higher \( RC/S_o \) scores. Overall, the best detector with considerable stable performance is KAZE detector, followed by SIFT detector in this study.

In our study, we further assess local feature detectors on six pairs of images (from Fig. 2(c)-2(h)) and present the results in Fig 5. It is clearly seen that both Harris and BRISK detectors barely detect enough correct matches and they spread over only a minor area of overlapping region with higher \( RC/S_o \) scores when compare to others. KAZE features exhibits good performance as it yields higher number of correct matches over a wider area of overlapping region with lower \( RC/S_o \) scores. Although SIFT and SURF detect lower number of correct matches compare to KAZE detectors, both are able spread the correct matches over a wider area of overlapping region that may well benefit in upsurge the stitching result. As we mentioned in Section I, the result of image stitching will be better when reliable corresponding keypoints are enough and wide-spread evenly over the whole area of overlapping region. Nevertheless, our study shows that none of the detectors can acquire more than 50% spread-overlap scores for all near-uniform scene image pairs (see Fig. 5(b)). Consequently, this would likely introduce misaligned, distortion or visible artefacts in image stitching.

**CONCLUSIONS**

This paper presented a new performance evaluation of local features using especially the near-uniform scene images based on repeatability and recall \( RC \) measure. Besides, we have also introduced a new evaluation criterion, that is spread-overlap \( S_o \) and \( RC/S_o \) scores, in which we suggest to use them in comparing the spread of correct matches over the area of overlapping region between near-uniform scene images. In all cases of our study, the KAZE detector performs better than other detectors in terms of detecting sufficient reliable features that spread over a wide area of overlapping region. But then again, none of the detectors can reach more than 50% spread-overlap scores for all inspected near-uniform scene images that may likely to introduce misaligned, distortion or visible artefacts in image stitching.

**ACKNOWLEDGMENT**

This work is supported by the Ministry of Education Malaysia through the provision of FRGS research grant, F02/FRGS/1492/2016.
REFERENCES


