A Performance Evaluation of Feature Descriptors for Image Stitching in Architectural Images

Prashanth Balasubramanian, Vinay Kumar Verma and Anurag Mittal

Computer Vision Lab,
Indian Institute of Technology Madras,
Chennai-600036,
India
{bprash,vkverma,amittal}@cse.iitm.ac.in

Abstract. We present a performance comparison of 4 feature descriptors for the task of feature matching in Panorama Stitching on images taken from architectural scenes and archaeological sites. Such scenes are generally characterized by structured objects that vary in their depth and large homogeneous regions. We test SIFT, LIOP, HRI and HRI-CSLTP on 4 different categories of images: well-structured with some depth variations, partially homogeneous with large depth variations, nearly homogeneous with a little amount of structural details and illumination-variant. These challenges test the distinctiveness and the intensity normalization schemes adopted by these descriptors. HRI-CSLTP and SIFT perform on par with each other and are better than the others on many of the test scenarios while LIOP performs well when the intensity changes are complex. The results of LIOP also show that the order computations of the pixels have to be made in a noise-resilient manner, especially in homogeneous regions.

1 Introduction

Identification of point-correspondences between images is an important problem that finds application in many tasks such as Registration, Stitching, Disparity Matching, 3-D Reconstruction, Tracking, Object Identification and Classification. As the transformations between the images are seldom known a priori, the practice is to localize on distinctive regions of images (called as *keypoints*) and match them under different transformations. Matching of keypoints across 2 images is done by building *feature descriptors* that express the visual characteristics of the regions around the keypoints, and correspond them using a suitable distance metric. The descriptors are expected to be sufficiently distinctive so as to represent the keypoint and be robust to geometric transformations, illumination variations, different blurs, artifacts due to sampling and compression.

Many interesting attempts have been made to design descriptors which satisfy these said characteristics. Early work used the raw pixels of the regions around the keypoints and studied their correlation measure. As correlation measures do not consider geometric information, such measures cannot tolerate localization errors of keypoints, and so are good when the regions are exactly registered. Further, these measures can only handle linear changes in intensities while it is well-known that non-linear variations in illuminations are commonplace occurrences, especially in the under-saturation and over-saturation regions.

Gradient-based methods have proposed effective strategies to handle many of these challenges. The popular SIFT[1] algorithm captures the local gradient distributions around the keypoints. Bay et.al [2] propose a faster variant of SIFT called as *SURF*, by computing Haar-wavelet responses using integral images. It is also compact(64 dimensions) and uses the sign of the Laplacian to perform faster indexing. The *GLOH* descriptor [3] improves the robustness and distinctiveness of *SIFT*. It divides the region into a log-polar network of 17 spatial bins, on each of which is a 16 dimensional orientation histogram built. PCA is used to reduce the 272 dimensions to 128 which are used in matching. Ke and Sukthankar [4] propose a dimensionally reduced descriptor *PCA-SIFT* by vectorizing the x and y gradients of the pixels of the normalized patch and linearly projecting the vectors onto a much lower-dimensional (~30) eigen-space. They argue that an eigen-projection is sufficient to model the variations in the 3D-scene and viewpoints, although the evaluation in [3] shows other descriptors to perform better. Shape Context[5] is another method that bins the orientations of pixels into a log-polar grid. Although the authors applied it only for edge point locations and not orientations, it can be used as a region descriptor as well [3]. Apart from these, there are also other modifications of gradient histograms such as those in [6,7,8].

Order-based descriptors that are constructed based on the sorting of pixels are an alternative strategy to gradient-based descriptors. Zabih & Woodfill[9] proposed two techniques - *rank* and *census transforms* - that are based on the order of intensities of neighbors of a pixel and the count of flipped point-pairs. Such order-based methods are inherently invariant to monotonic changes in illumination. However, they fail in the presence of pixel noise as a single salt-and-pepper flip can change the counts, which is alleviated to a certain extent by Bhat & Nayar [10]. Mittal & Ramesh [11] improve the latter by penalizing an order-flip in proportion to the change in the intensities of the pixels that underwent the flip. This helps to prevent the movement of pixels due to Gaussian noise. Tang et.al [12] propose the *OSID* descriptor that builds a histogram of orders computed on the entire patch. Though invariant to monotonic illumination variations, it can fail on a patch having many pixels of similar intensities as these tend to shift under Gaussian noise. Gupta & Mittal [13] alleviate this problem by designing a histogram of relative intensities whose bins are adaptively designed for the saturated and the non-saturated regions. Wang et.al [14] improve upon this in their *LIOP* descriptor by inducing rotation invariance to it. The motivation is based on their study [15] that identifies estimation of keypoint orientation as a major source of localization error.

There are other variants of order-based descriptors that are bit-strings of comparisons of pixels. These are attractive because of their minimal storage requirements and their ability to be compared fast. *Local Binary Patterns(LBP)* [16], first applied for face recognition and texture classification, are formed by the comparison of a pixel with its neighbors and constructing a histogram of these patterns. Since these patterns are rather high-dimensional, variants such as [17,13] compare only certain pixels in the neighborhood without sacrificing the discriminative ability of the *LBP* patterns. Calonder et al. [18] propose the *BRIEF* descriptor that randomly samples 128 or 256 pixel-pairs from the smoothed patch and forms a bit-string based on the outputs of their comparisons.

The bit-string turns out to be, surprisingly, discriminative. Because of the manner in which it is constructed, *BRIEF* is not rotation-invariant and Rublee et al.[19] propose the *ORB* descriptor that makes *BRIEF* rotation invariant. Leutenegger et al.[20] design a variant of *BRIEF* called as *BRISK* [20] that is formed by the comparisons of pixels placed uniformly on concentric circles. The region is rotation-normalized according to the orientation estimated from the pixels on the circles. To avoid aliasing while sampling points from the circles, each point is smoothed by a Gaussian window of width that is sufficient to not distort the information content of close-by points. They also propose a fast keypoint detector. The *FREAK* descriptor by Alahi et al. [21], is another binary descriptor that compares intensities of pixels sampled in a pattern as observed in the human retinal system. They also outline the reason behind why such comparison-based binary descriptors work, based on studies of the human visual system.

Mikolajczyk and Schmid [3] provide an extensive comparison of many keypoint descriptors including *SIFT, SURF, Shape-Context, SIFT-PCA, GLOH, Cross-correlation and Steerable Filters* and observe that, although *SIFT* performs well in many scenarios, there is no one particular descriptor which works for all cases. A comparison of the modern descriptors has been made independently by Miksik & Mikolajczyk [22] and Heinly et al. [23].

In this paper, we aim to study the performance of 4 descriptors - *SIFT*, *LIOP*, *HRI* and *HRI-CSLTP* for matching keypoints in the applications of Stitching of images of architectural scenes. Such images are characterized by well-structured and textured monuments that can be varying in depth, may have large areas of homogeneous regions especially when shot for a panoramic mosaic and can have varying illumination levels. Accordingly, we test these descriptors on 4 kinds of images from a dataset of archaeological sites and historical monuments: $1)$ well-structured with sufficient depth variation $2)$ partly structured and partly homogeneous $3)$ nearly homogeneous with a few structured regions and $4)$ illumination change on a dataset . We aim to study the scope of application of these descriptors by testing them on the said challenges. To that end, we plot their response graphs for matching, compare their performance and draw conclusions therefrom.

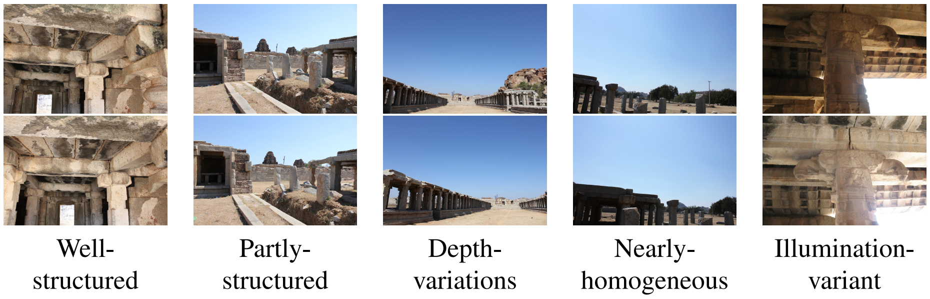
The paper is organized as follows. Section 2 briefly discusses the challenges that are usually posed by architectural scenes with visual examples. An overview of the descriptors that are tested in the paper is given in Section 3. Section 4 discusses the dataset, the groundtruthing technique, the evaluation methodology adopted to test the descriptors. The experimental results are presented along with their analyses in Section 5. Section 6 concludes the paper with the lessons drawn from the experiments.

2 Architectural scenes and their challenges

Fig. 1 shows some images from a typical dataset of archaeological sites and historical monuments. Such monuments are usually structured [1] with repeated occurrences of textured regions (col.1 of Fig.1) at varying levels of depths (cols. 2 & 3 of Fig.1).

[1] A region of image is well-structured when it is characterized by regular occurrences of homogeneous or textured patches that are flanked by well-defined object gradients. A typical example is that of a building, as opposed to an image of a scenery.

These images may also include large homogeneous regions, especially when shot for a panoramic mosaic or 3-D reconstruction, with a vacant landscape in the front or sky in the back (col.4 of Fig.1). Homogeneous regions are poor conveyors of distinctive visual information. So, when large areas of images are covered by homogeneous regions, it becomes important to match the available keypoints from the non-homogeneous regions in a reliable and correct manner, and discard as many pseudo-matches as possible. The descriptors have to be highly distinctive to suit this requirement. Further, the lighting conditions and the time of the day when the images are shot govern the intensities of the pixels and can make them vary in a non-linear way (col.5 of Fig.1), especially in under-exposed or over-exposed regions (for instance, interior structures that are poorly lit). The descriptors need to be resilient to these changes in intensities by adopting a generic normalization technique.



Well-structured

Partly-structured

Depth-variations

Nearly-homogeneous

Illumination-variant

Fig. 1. Challenges that usually beset a feature matcher.

In the next section, we present a brief overview of the 4 descriptors - *SIFT, HRI, HRI-CSLTP, LIOP* - that are tested in these challenges. While *SIFT* [1] is well-known, *HRI-CSLTP* [13] and *LIOP* [14] are recent order-based descriptors that have performed well on the standard datasets [13,14,22].

3 Overview of the Descriptors

SIFT descriptors [1] capture the local distribution of the gradients in the patches around the keypoints by tri-linearly binning the gradient magnitudes of the pixels into 8 orientation bins. To make the descriptor robust to small pixel-movements, a patch is divided into 4x4 spatial-grids over which the orientation histograms are built which are then concatenated to form the descriptor of the patch. Robustness to spikes in gradient-magnitudes is handled by capping the gradient-magnitudes to be a maximum of 0.2 and $l2$ normalization of the descriptors make brings resilience to linear changes in illumination. Each patch yields a 128 dimensional, real-valued descriptor.

HRI descriptors [13] capture the relative orders of the pixels of the patch based on their intensities. Orders have the natural ability to be invariant to monotonic changes in illumination. In a *HRI* descriptor, pixels bin their intensities into intervals that are designed based on the intensity distribution of the overall patch. Linear normalization of intensities yields illumination invariance, wherein the min and the max points of the normalization are adaptively chosen for the saturated and the non-saturated regions[2]. Gaussian pixel noise is handled by a uniform distribution of the intensities into the intervals, and trilinear interpolation and spatial-division of the patch into grids handle small pixel movements. It is to be noted that gradient information is not used, in contrast to SIFT [1].

CSLTP descriptors [13] look at the intensity differences of the diagonal neighbors of each pixel and encodes them using 3 categories based on a threshold parameter, T; two of the categories identify differences of opposing contrast, $|i_1 - i_2| > T$, while the third identifies pixels of nearly equal intensities, $|i_1 - i_2| \leq T$. T helps to choose a certain amount of separation between the diagonal pairs. With 2 diagonal pairs, each being encoded with 3 patterns, there are totally 9 different neighborhood patterns which can be treated as the 9 bins of the *CSLTP* histogram. Based on its pattern, each pixel contributes a weighted vote to one of the 9 bins. The weight is designed to eliminate a pixel if it has nearly homogeneous neighbors and, thereby, prevent its movement. The patch is divided into 4x4 grids to counter small spatial errors and the *CSLTP* histograms of the grid are concatenated to yield the *CSLTP* descriptor of the patch.

LIOP descriptors [14] are designed to be rotation and monotonic-illumination invariant by using the order of the intensities of the pixels. The local intensity order pattern of a pixel is a weighted vector that encodes the ranking of its 4 neighbors. The neighbors are sampled from a circular neighborhood in a rotation-invariant manner to avoid the errors in estimation of keypoint orientation [15]. Gaussian noise is handled by giving more weights to the patterns that result from neighbors differing in their intensities by a certain threshold. In addition to the local patterns, the patch is intensity-thresholded using multiple values to yield regions of similar intensities, called as ordinal bins. The *LIOP* pattern of an ordinal bin is the weighted summation of those of its pixels; these *LIOP* patterns are concatenated in the order of the ordinal bins resulting in a rotation-invariant *LIOP* descriptor of the patch.

4 Dataset & Evalution Criterion

We evaluate the descriptors on an architectural dataset which contains images of many archaeological monuments and historical sites. The images, $\sim$ 50K in all, have been shot in two resolutions (1280 x 960 & 3648 x 2736) and are categorized according to varying details of the structures of the sites and thus, made suitable, for different tasks such as panorama stitching and 3D-reconstruction.

For testing the descriptors on image registration for Mosaicking, images shot with the panoramic constraints[3] have been chosen. Following are the challenges based on the nature of the scene that have been used to test the descriptors: 1) well-structured with sufficient depth variations 2) partly structured and partly homogeneous 3) nearly homogeneous with a few structured regions and 4) illumination changes. Estimation of

[2] A region is saturated if its pixels have intensities either below 10 or above 245.

[3] A set of images is suitable for panoramic stitching if all of them depict a planar scene or are shot with the camera center being fixed.

homography for a pair of images is done with the manual input of 4 point correspondences.

We use the evaluation criterion proposed by Mikolajczyk and Schmid [3] that identifies the correct and the false descriptor matches using ground truth correspondences at a particular region overlap error(50% in our experiments), as defined by Mikolajczyk et al. [24]. The descriptor matches are obtained using the ratio-test proposed by Lowe [1], the threshold for which is varied to obtain the points on the Precision-Recall response graphs. The correspondences of the regions for a particular overlap error (50%) and the validation of the descriptor matches have been computed using the code available at the Affine Covariant Features page[4].

DoG keypoints [1] are detected using the co-variant feature detector routine in the *VL-FEAT* library [25]. The minimum absolute value of the cornerness measure is empirically set to 3 for all the experiments. For the *SIFT* and the *LIOP* descriptors, the implementations in the *VL-FEAT* library are used. *HRI* and *HRI-CSLTP* have been implemented by us.

5 Performance Evaluation

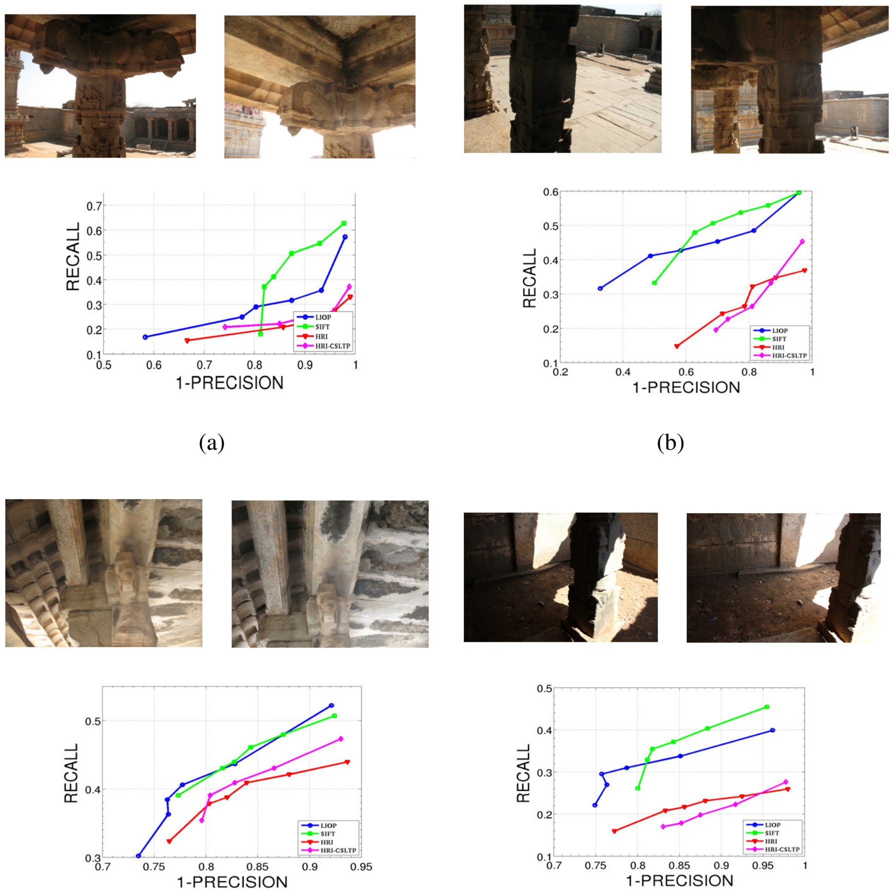
5.1 Images with Illumination variations

Images taken in an uncontrolled environment such as archaeological sites exhibit wide variety of intensity ranges depending on the ambient light which need not illumine the objects in the scene uniformly, especially the interior parts of structures and can thus, result in under-saturated or over-saturated regions. Such variations in the intensities are usually non-linear and hence, the descriptors have to deal with an appropriate normalization scheme. Fig.2 shows the performance of the descriptors on images that vary in their illumination patterns. These are usually indoors where the natural light doesn't reach all portions of the scene uniformly. The recall rate is generally low as it is ~30% when the precision is $\sim$ 30% for the best performer(s), except in Fig.2(b) which might be due to the good matches from the well-lit outdoor structures. *SIFT* seems to be doing consistently well,although *LIOP* is not far behind. Though *HRI-CSLTP* and *HRI* use adaptive binning, the changes in these images might be very non-linear for these methods to perform well.

5.2 Structured Images

Fig.3 shows the performance of the descriptors for images that are well-structured with some depth variations and nearly well-lit light conditions. The aim here is to study if the descriptors can match the keypoints output by the detector when they vary in their texture content due to depth and viewpoint changes. *SIFT* and *HRI-CSLTP* perform consistently well in all the 4 cases. The additional edge direction information in *HRI-CSLTP* definitely helps it score better than *HRI*, although the marginal differences in their performances might suggest that *CSLTP* may have to be combined with other descriptors as it captures directional information only in 4 orientations.

[4] http://www.robots.ox.ac.uk/~vgg/research/affine/desc_evaluation.html#code



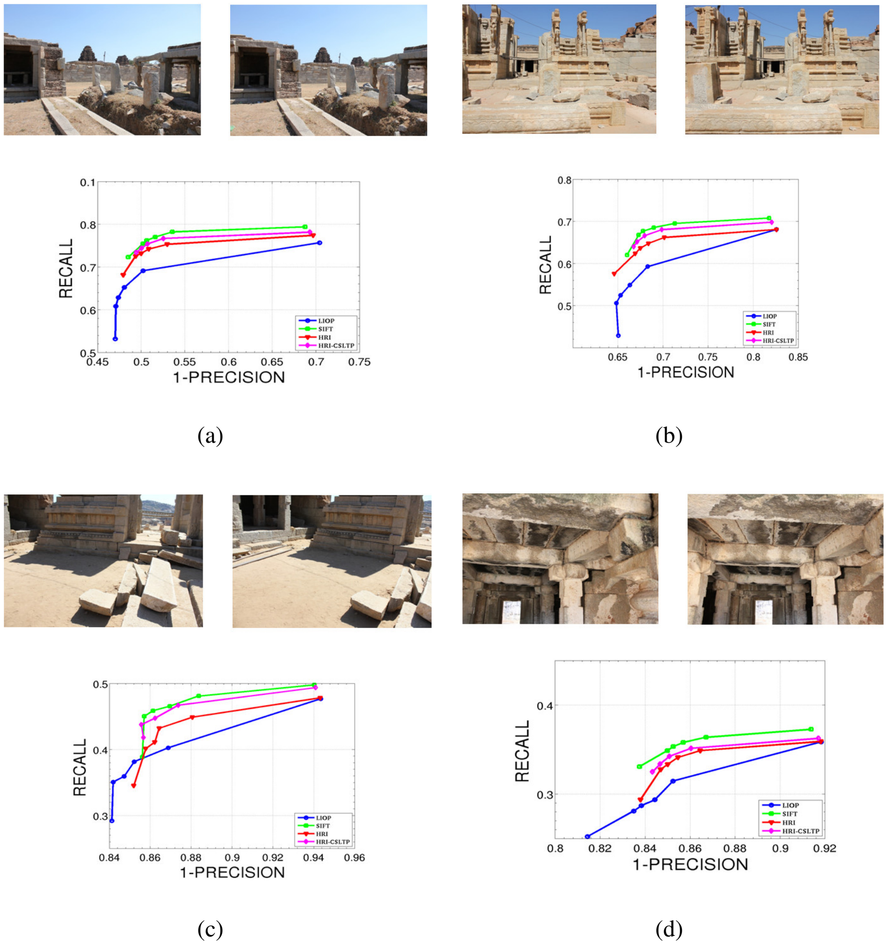
(a)

(b)

(c)

(d)

Fig. 2. The performance of the descriptors on images with intensity variations. The ranges of the plots have been set different for the sake of clarity.



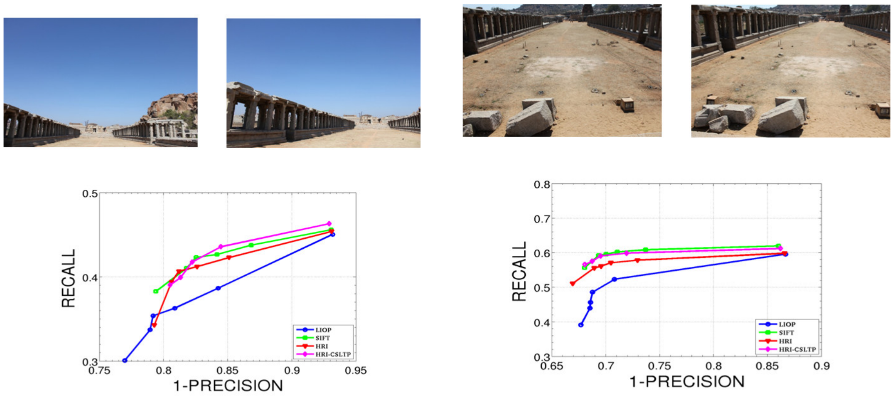
(a) (b)

(c) (d)

Fig. 3. The performance of the descriptors on well-structured images with some depth variations. The ranges of the plots have been set different for the sake of clarity.

5.3 Partially Homogeneous Images

Fig.4 shows the performance of the descriptors for images that are partially homogeneous containing large depth variations. Such images are usually captured to get a profile of the entire scene when it contains objects that vary significantly in their depths (e.g. a long wall flanked by a bare landscape on its side). For matching, the descriptors have to rely on the keypoints generated from the structured regions of the images. We find that *SIFT* and *HRI-CSLTP* perform well with the differences being very marginal in both the test cases. The orders of the pixels considered in *LIOP* can become noisy in homogeneous regions and that may explain the nature of its performance in these cases.

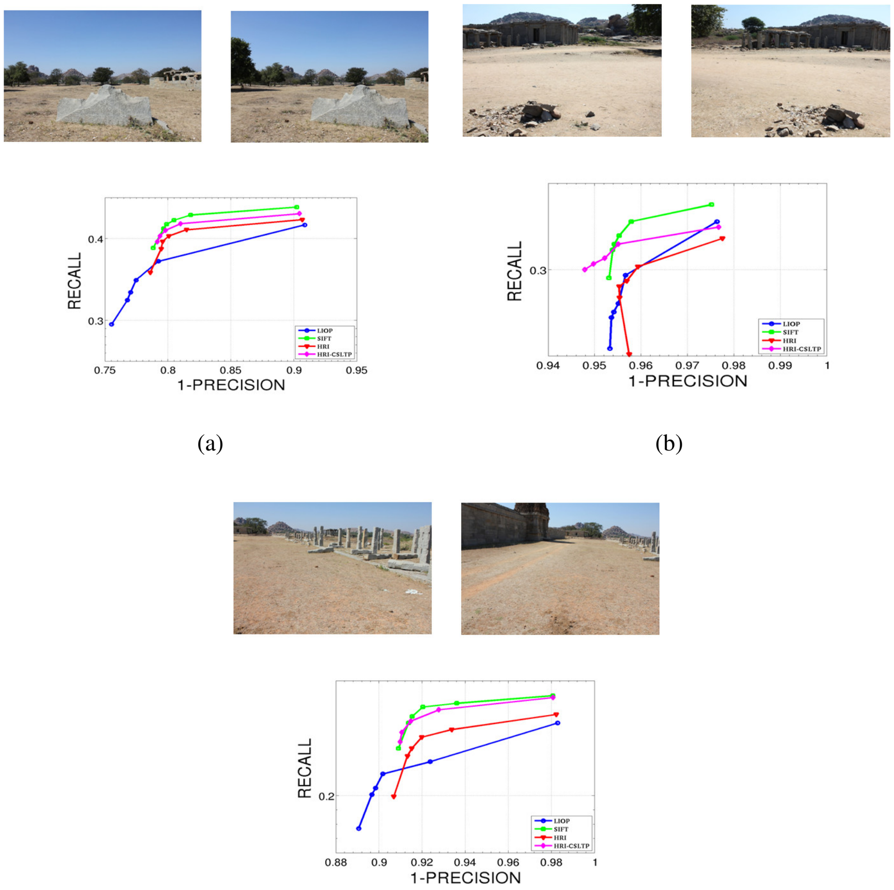


(a) (b)

Fig. 4. The performance of the descriptors on partially homogeneous images with significant depth variations. The ranges of the plots have been set different for the sake of clarity.

5.4 Nearly Homogeneous Images

Fig.5 shows the performance of the descriptors for images that are nearly homogeneous with very little amount of structures in them. Such images are usually captured in a panoramic shot of an architectural monument that has a nearly empty landscape in the front. The low ranges of precision in Fig.5 can be explained by the fact that nearly homogeneous regions tend to result in large number of false matches. The trend exhibited by the descriptors is the same as in the previous 2 challenges. Though the order patterns used in *LIOP* are weighted, the results suggest that the weighting might not be sufficient when there are large areas of homogeneous regions.



(a) (b)

(c)

Fig. 5. The performance of the descriptors on nearly homogeneous images with very little structures. The ranges of the plots have been set different for the sake of clarity.

6 Conclusions

We presented a performance evaluation of 4 feature descriptors for the task of feature matching in Image Stitching when the images are of archaeological scenes and architectural sites. As these images are characterized by structures that vary in their textural content and depth and homogeneous regions, we categorized the dataset into 4 classes and tested the descriptors on them. *SIFT* and *HRI-CSLTP* perform better than the others in many of the test cases highlighting their distinctiveness in representing the keypoint regions. *LIOP* performs well when the intensity variations are complex. Also, the results of *LIOP* show that the order computations have to be done in a noise-resilient manner, especially when homogeneous regions are involved. This performance evaluation can be extended to other applications like 3-D reconstruction to understand the scope of applicability of these descriptors.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
2. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In Leonardis, A., Bischof, H., Pinz, A., eds.: The proceedings of the 9th European Conference on Computer Vision. Volume 3951 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2006) 404–417
3. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1615–1630
4. Ke, Y., Sukthankar, R.: Pca-sift: a more distinctive representation for local image descriptors. In: The proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2004) II–506–II–513 Vol.2
5. Mori, G., Belongie, S., Malik, J.: Efficient shape matching using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1832–1837
6. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1265–1278
7. Mikolajczyk, K., Matas, J.: Improving descriptors for fast tree matching by optimal linear projection. In: The Proceedings of the Eleventh IEEE International Conference on Computer Vision. (2007) 1–8
8. Freeman, W., Adelson, E.: The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991) 891–906
9. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In Eklundh, J.O., ed.: The proceedings of the 3rd European Conference on Computer Vision. Volume 801 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (1994) 151–158
10. Bhat, D.N., Nayar, S.K.: Ordinal measures for image correspondence. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 415–423
11. Mittal, A., Ramesh, V.: An intensity-augmented ordinal measure for visual correspondence. In: The proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 1. (2006) 849–856
12. Tang, F., Lim, S.H., Chang, N., Tao, H.: A novel feature descriptor invariant to complex brightness changes. In: The proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2009) 2631–2638

13. Gupta, R., Patil, H., Mittal, A.: Robust order-based methods for feature description. In: The proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2010) 334–351
14. Wang, Z., Fan, B., Wu, F.: Local intensity order pattern for feature description. In: The Proceedings of the Thirteenth IEEE International Conference on Computer Vision. (2011) 603–610
15. Fan, B., Wu, F., Hu, Z.: Rotationally invariant descriptors using intensity order pooling. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 2031–2045
16. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 971–987
17. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recognition **42** (2009) 425–436
18. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: Brief: Computing a local binary descriptor very fast. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1281–1298
19. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: an efficient alternative to sift or surf. In: The Proceedings of the Thirteenth IEEE International Conference on Computer Vision, IEEE (2011) 2564–2571
20. Leutenegger, S., Chli, M., Siegwart, R.: BRISK: Binary robust invariant scalable keypoints. In: The Proceedings of the Thirteenth IEEE International Conference on Computer Vision. (2011)
21. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: The proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2012) 510–517
22. Miksik, O., Mikolajczyk, K.: Evaluation of local detectors and descriptors for fast feature matching. In: The 21st International Conference on Pattern Recognition. (2012) 2681–2684
23. Heinly, J., Dunn, E., Frahm, J.M.: Comparative evaluation of binary features. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: Computer Vision ECCV 2012. Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 759–773
24. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision **65** (2005) 43–72
25. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/` (2008)