Robust Order-based Methods for Feature Description

Raj Gupta, Harshal Patil and Anurag Mittal
Department of Computer Science and Engg.
Indian Institute of Technology Madras, Chennai INDIA - 600036
{rgupta,harshal,amittal}@cse.iitm.ac.in

Abstract

Feature-based methods have found increasing use in many applications such as object recognition, 3D reconstruction and mosaicing. In this paper, we focus on the problem of matching such features. While a histogram-of-gradients type methods such as SIFT, GLOH and Shape Context are currently popular, several papers have suggested using orders of pixels rather than raw intensities and shown improved results for some applications. The papers suggest two different techniques for doing so: (1) A Histogram of Relative Orders in the Patch and (2) A Histogram of LBP codes. While these methods have shown good performance, they neglect the fact that the orders can be quite noisy in the presence of Gaussian noise. In this paper, we propose changes to these approaches to make them robust to Gaussian noise. We also show how the descriptors can be matched using recently developed more advanced techniques to obtain better matching performance. Finally, we show that the two methods have complimentary strengths and that by combining the two descriptors, one obtains much better results than either of them considered separately. The results are shown on the standard 2D Oxford and the 3D Caltech datasets.

1. Introduction

The use of features for image representation and matching has gained tremendous importance and popularity in recent years for problems as diverse as Image Alignment, mosiacing, 3D Reconstruction, Object Recognition and Tracking. Features are extracted using methods such as Harris features[4], Harris-affine, Hessian, Hessian-affine[13], MSER (Maximally Stable Extremal Regions)[20], DOG (Difference of Gaussians)[10] and others[6, 25], methods for matching (normalized) patches include the popular SIFT (Scale Invariant Feature Transform)[10] and its variants such as GLOH (Gradient Location and Orientation Histogram)[15], Shape Context[18] and other modifications of such gradient/edge based methods such as [1, 7, 8, 18, 12].

More recently, methods have been proposed[11, 3] that use orders of pixels rather than raw intensities and the results from these methods are encouraging. In this paper, we propose new ways of using the order between pixels that are more robust to noise in the underlying data. The two methods proposed capture orthogonal properties of a feature region - one captures the overall distribution of pixels in the patch and the other captures local gradient properties.

The first method is based on orders of pixels relative to the entire patch and builds a histogram based on the relative position of the intensities w.r.t. to the entire patch. The second method looks at local orders of pixels and generalizes the Center-Symmetric Local Binary Patterns (CS-LBP) descriptor. Specifically, instead of a binary code, we develop a ternary code, which we call Center Symmetric Local Ternary Patterns (CS-LTP). Both these methods are designed to be more robust to Gaussian noise than previously considered descriptors. They capture orthogonal information and a combination of these two methods was found to improve upon either of the two considered separately.

2. Related Work

The idea of matching images/features using order of intensities rather than raw intensities is not new. By considering only the orders between pixels rather than their intensities, one obtains invariance to a monotonic change in the intensities. The Census algorithm[26] transforms the intensity space to an "order" space, where a bit pattern is formed by looking at the orders of a given pixel with its neighbors. This algorithm essentially counts the number of flipped point pairs in the patch. Bhat and Nayar[2] use an improved version of this algorithm where they somewhat alleviate the problem of counting even one salt-and-pepper error in a pixel multiple times. Mittal and Ramesh [16] proposed a method in which the penalty for an order flip is proportional to the intensity difference between the two flipped pixels. This reduces the error due to pixels whose order may have got flipped due to Gaussian noise. Finally, Singh et al [23] present a statistical approach whose match measure

can be tuned to the underlying error process. All of these methods assume that the pixel locations don't vary across the two patches and are thus inappropriate for the feature matching problem where the pixel locations might undergo some shift.

The LBP Descriptor, which is based on relative order of neighboring pixels has also shown promise for several applications. Binary Patterns are created for each pixel by comparing a pixel value with its neighboring intensities. The histogram of the such binary patterns computed over a region is used for texture description in [19]. As the LBP operator produces a rather higher dimensional histogram and is therefore difficult to use in the context of a region descriptor, a Center-Symmetric LBP which only compares center-symmetric pairs of pixels (Fig. 2) was considered for feature description in [5]. Recently, there have also been papers that develop a descriptor based on the overall order of the pixels in a patch. [11] have proposed converting any descriptor to the order space by simply forming the descriptor in the normal way (i.e. by using gradient information) and then considering the ordinal information of the descriptor values. [3] have proposed building a histogram of orders, where the orders are computed with respect to the entire patch.

Both these methods for feature description use only orders and completely neglect the intensities. While this gives invariance to monotonic change, the orders can be noisy in the presence of Gaussian noise, especially when the nearby pixels are close in intensity. In this paper, we propose methods that are more robust to Gaussian noise, although they are still based on orders. Experiments show improved performance over standard datasets.

3. Histogram of Relative Intensities

The basic idea is to use the intensity directly rather than gradients. In order to obtain invariance to illumination changes, the range of the intensity values is first determined, which is used to normalize the intensities. While the smallest and largest pixels can be used, we make it more robust by using the average of the first j and last j pixels for the normalization (j of around $1/32$ of the total number of pixels was found to give the best results). Since our intensity normalization assumes a linear change of intensities and a non-linear effect takes place due to under-saturation and over-saturation, these values can be noisy when these values are close to 0 and 255. Thus, we employ an adaptive scheme whereby we use the lowest block of pixels that give a value above 10 and the highest block of pixels that give a value below 245 and use these to normalize the intensities. In doing so, we assume a uniform distribution of the pixels (note that the range may sometimes go out of 0-255 when we do this). We also tried using simply the mean and standard deviation for this normalization, but the results were worse than the method used.

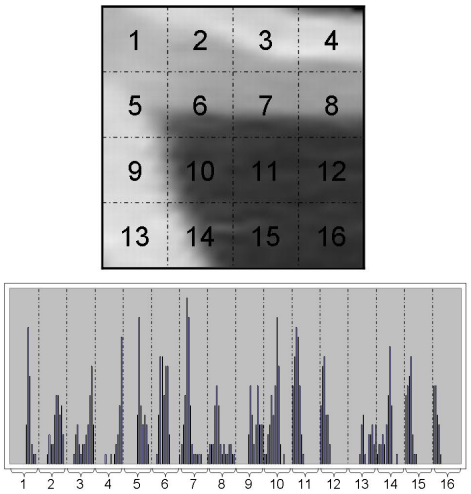


Figure 1. Illustrative histogram for our descriptor for the patch shown on the top.

Once we determine the starting and ending point of the intensity range, we obtain intervals by dividing this range into k equal intervals based on intensities. Note that this is different from the earlier proposed method which we call Histogram of Orders (HOO)[3] in that HOO forms intervals based on orders and we form intervals based on intensities. Our method is more robust to Gaussian noise since small changes in intensities due to noise does not lead to large changes in the descriptor whereas if a lot of points have similar intensities in a patch, then even small changes in the intensites can lead to a large change in the order of the pixels.

Now, the patch is divided into $s \times s$ spatial bins in a manner similar to the SIFT descriptor and at each spatial bin, k bins are created where the j-th bin stores the number of pixels in that spatial region that have their intensities in the j-th interval as determined above. Similar to the SIFT descriptor; we distribute the weight of each pixel into adjacent histogram bins using trilinear interpolation depending on the exact intensity and location of the pixel. This helps in gaining robustness to small localization and normalization errors common in the feature extraction and normalization process. Furthermore, similar to the SIFT descriptor, we also give more weight to the center pixels as opposed to the boundary ones as these pixels are more stable and more likely to be always present in a corresponding patch. The σ for this Gaussian weight function is set to half of the spatial descriptor window size.

We thus have a total of $s \times s \times k$ bins in the descriptor. This is shown diagrammatically in Fig. 1 for the image patch shown in Fig. 1. Best results were obtained for a descriptor of size $4 \times 4 \times 16 = 256$ and this value was used

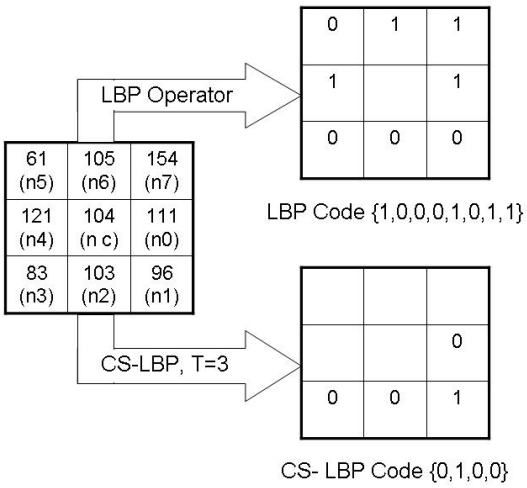


Figure 2. Illustrative diagram for CS-LBP Operator

for all the results in the paper. We call this method HRI (Histogram of Relative Intensities).

4. Center Symmetric Local Ternary Patterns

The above-mentioned approach works on the overall distribution/order of the pixels in the patch but does not capture local gradient information. Such information can be useful since it is orthogonal to the global order information in the patch. In the second part of our descriptor, we propose a method that works on the local gradient information. In this method, we construct a histogram of Center-Symmetric Local Ternary Patterns(CS-LTP) accumulated in spatial bins similar to the previous approach. The CS-LTP codes are a variant of the CS-LBP codes proposed in [5] and were found to give superior performance to CS-LBP codes in most of our experiments.

4.1. Center-Symmetric-Local Ternary Patterns

The CS-LBP descriptor has recently been proposed by [5]. In these descriptor, the LBP operator has been modified such that at each pixel, neighboring pixels that are opposite to each other are compared in order to generate a binary code. This is illustrated in Fig. 2. Since only 4 comparisons are made, we get histograms of size 16 at each spatial bin. In this work, we have modified the CS-LBP descriptor in several ways in order to improve upon the performance of this descriptor. First, since the order of pixels in homogenous regions is very noisy, we propose using a third value which states that the orders of two pixels are almost the same, i.e. within some threshold value. This gives us ternary codes (Fig. 3). However, if we were to use 4 comparisons as in CS-LBP, we would get a histogram of size 81.

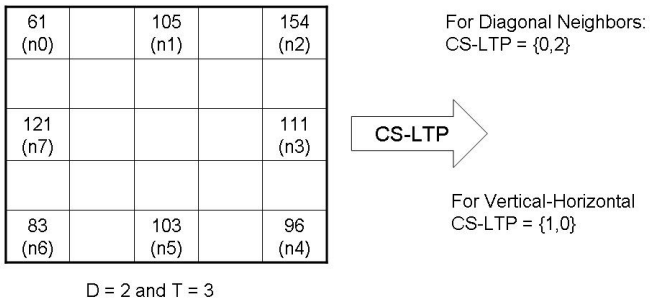


Figure 3. Illustrative diagram for CS-LTP Operator

In order to reduce the size of the histograms, we only consider two comparisons as shown in Fig. 3. Due to feature normalization (that typically puts high gradients along the x or y axis) and general image characteristics, we found that using only the diagonal comparisons to generate the CS-LTP code for each pixel in the patch was sufficient while the vertical and horizontal comparisons were quite noisy. Using only these two comparisons, we obtain a histogram of 9 bins for each spatial bin..

Mathematically, the Center-Symmetric Local Ternary Pattern at point p with center symmetric pairs of pixels at a d_n neighboring distance is given by (Fig. 3):

$$CS-LTP(p,d_n,T) = f(n_0-n_4)+f(n_2-n_6)\times 3 \quad (1)$$

where

$$f(x) = \begin{cases} 0 & x < -T \\ 2 & x > T \\ 1 & \text{else} \end{cases}$$

For our experiments, we have chosen $d_n = 2$ and $T = 3$.

As in previous methods, a bilinear interpolation is used to distribute the weight of each feature into adjacent bins. Also, it was found that the codes corresponding to 1, i.e. ones that say that the two matched points are almost the same, were less reliable than the other ones and had more tendency to shift. Because of this, we give less weight to the bins that correspond to codes having a 1. In particular, the weight for a bin corresponding to code (t_1, t_2) is taken to be $|t_1 - 1| + |t_2 - 1|$. We found much improved performance with this modification. It is to be noted that under this weighting scheme, the code 11 receives zero weight, i.e. the homogenous regions are totally neglected in such a scheme. Thus, the number of bins at each spatial bin is further reduced to 8, yielding a CS-LTP feature descriptor of size $4 \times 4 \times 8 = 128$. It may be noted that the CS-LTP codes we develop are different from the LTP codes proposed by [24] in that [24] separate the positive and negative values into two codes (i.e. use two thresholds of $\pm\delta$ to obtain two

codes) while we use the two thresholds together to obtain a ternary code. We believe this is more robust, although it can increase the descriptor size if not used carefully.

The two descriptors developed by us give orthogonal information. While the first one gives information about the overall distribution of intensities the patch, the second one encodes local gradient information. These two may be concatenated for improved results. We call this concatenated descriptor the HRI-CSLTP descriptor in the rest of the paper. The total size of this descriptor is $256 + 128 = 384$.

5. Descriptor Matching

The two histograms obtained by our methods may be matched using the common bin-by-bin L_2 or L_1 distance measures with fairly good results. At the same time, the Earth Movers Distance(EMD) has shown to give superior performance for many descriptors[9, 21] as it can account for possible shifts of values to nearby bins.

The Earth Mover's Distance (EMD) is measure of distance between two distribution d_1 and d_2 over some region R. It denotes, the minimum cost required to convert distribution d_1 to distribution d_2. The EMD generally defined over two distributions having same integral for given region such as normalized histogram. The EMD computation is based on solution of Hitchcock transportation problem, where one distribution d_1 is considered as supplier and distribution d_2 as consumer and cost associated is nothing but the distance between an element of d_1 and an element of d_2. Intuitively, the cost is measure of minimum amount of work needed to remove dissimilarity between two distributions. In case of histograms, the EMD [22] is defined as the minimal cost that must be paid to transform one histogram into the other, where there is a *"ground distance"* between the basic features that are aggregated into the histogram. Given two histograms P, Q the EMD as defined by Rubner et al. [22] is:

$$EMD(P,Q) = \min\{f_{ij}\}\frac{\Sigma_{i,j}(f_{i,j}d_{i,j})}{\Sigma_{i,j}f_{i,j}} \qquad (2)$$

where each f_{ij} represents the amount transported from the i^{th} source histogram bin to the j^{th} destination histogram bin. The distance d_{ij} the ground distance between bin i and bin j in the histograms.

In particular, [21] have recently shown that their usage of the EMD matching technique gives a small improvement in the matching accuracy of SIFT while many other ways of using EMD such as the L_1 or L_2 "ground distance" based one in fact reduces the discriminability of the descriptor, at least for the standard dataset from Oxford[14]. They connect only the adjacent bins in the gradient orientation dimension and do not do so in the spatial dimension. Furthermore, they only connect the adjacent bins with a cost of 1 and give a fixed cost of 2 for movement of more than 1 bins, considering such movement as being due to outliers. The matching method uses max flow customized for this problem and is much faster compared to other methods of computing EMD.

We have found that using the approach of [21] gives improvement for our descriptor as well. Therefore, we show the results in this paper using both the bin-by-bin L_2 distance and the EMD matcher of [21]. Similar to the results for SIFT as found by [21], other methods for using the EMD reduced the matching accuracy. For our usage of the EMD, we connect the adjacent bins in the order dimension as there might be some movement across these bins due to some extra or missing pixels in one of the patches compared to the other. The circular property i.e. possible movement from last bin to the first bin, holds only for SIFT and CS-LTP histograms and not for HRI and for HRI, we do not connect the arcs that make this EMD matcher circular. Furthermore, the different weights of CS-LTP bins is not a good choice while using the EMD matcher since a given pixel must be given the same weight for all bins. Thus, no weighting is used for CS-LTP bins while using EMD. However, the code 11 is again neglected and the other codes are considered in a circular fashion. This was found to give good results. However, since no such simple (single) adjacency is possible for CS-LBP codes, the EMD matcher was not used for CS-LBP and the default L_2 distance was used for comparison purposes.

6. Experiments and Results

We demonstrate the results of our experiments on two datasets: the '2D' Oxford dataset which tests robustness to different image degradations in images such as illumination changes, blur, JPEG compression, zoom, rotation changes and affine/viewpoint change and the '3D' dataset from Caltech which tests the distinctiveness of feature description on 3D objects.

6.1. The '2D' Oxford dataset

For the proposed descriptor, we first compare our results on the standard dataset obtainable from Oxford university site *http://www.robots.ox.ac.uk/ṽgg/research/affine*. Although many descriptors exists, for clarity purposes, we compare results on this dataset only with SIFT, HOO[3] and CS-LBP [5] since the first one was shown to be among the best in a standard evaluation [15] uptil a couple of years back and the approaches of HOO and CS-LBP are close to our approach. Circular binning can also be used with our method as in GLOH [15] and [3]. The relative performance of the descriptors is very similar to the experiments on square bins that we show in this paper.

The dataset contains images with different geometric

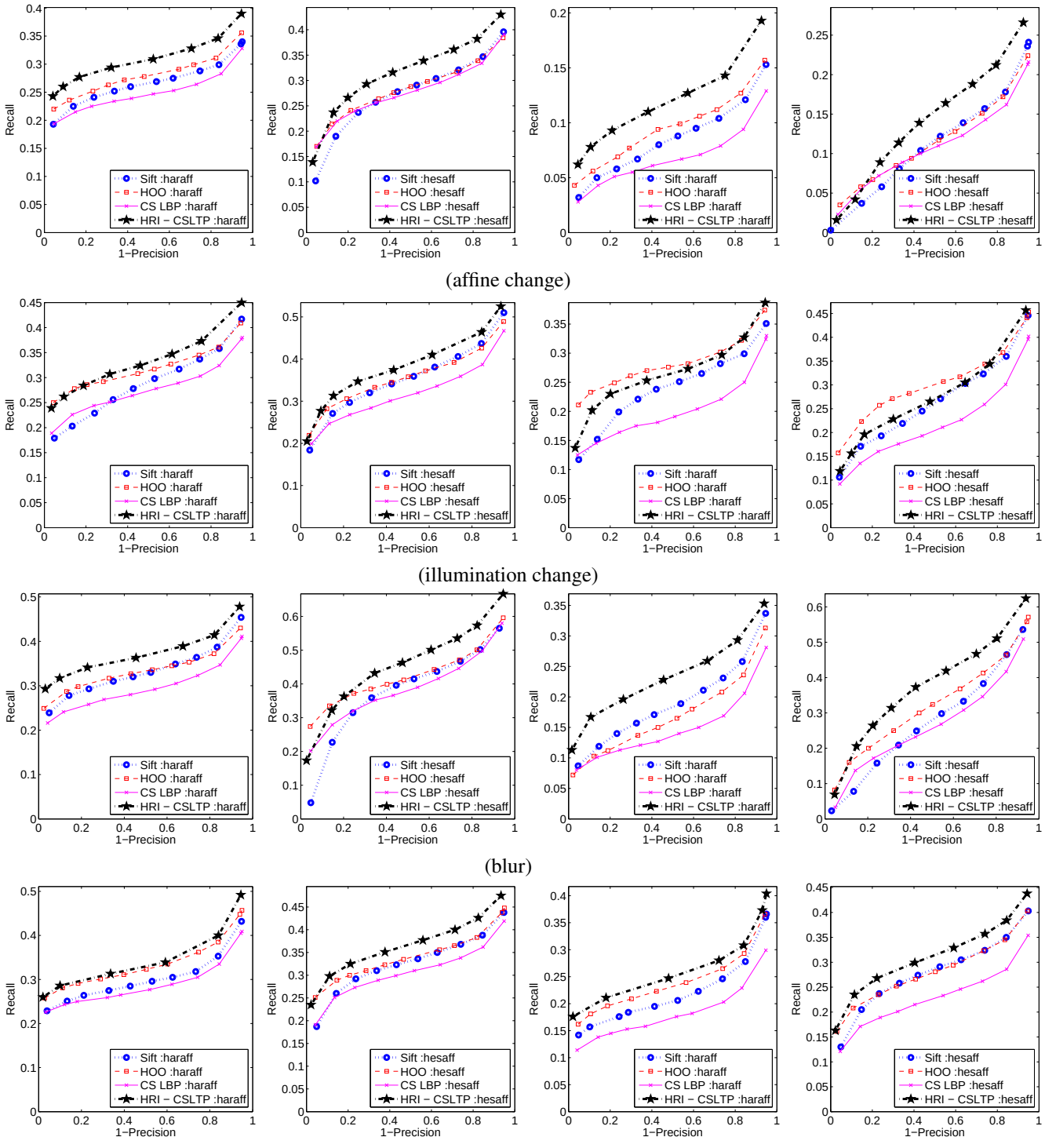
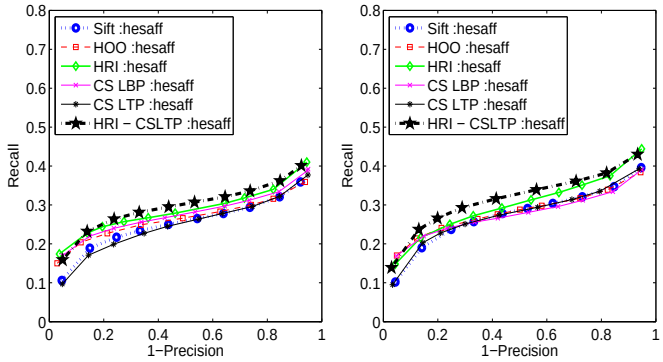
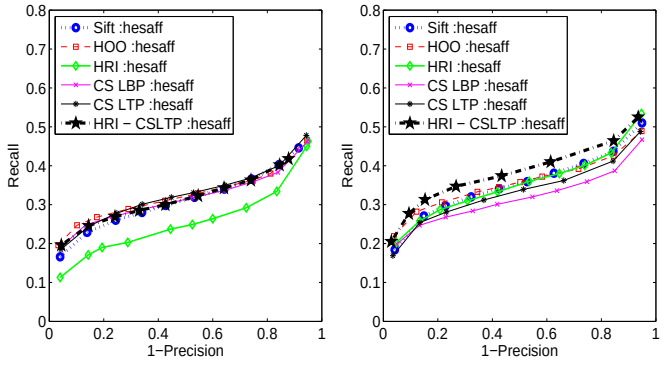


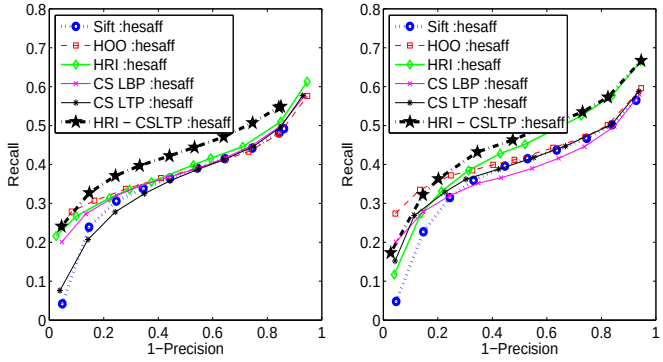
Figure 4. Comparison of SIFT[10], Histogram of Orders(HOO)[3], CS-LBP [5], and HRI + CS-LTP (proposed method) on (i) affine changes (graf), (ii) illumination change (leuven), (iii) blur (bikes) and (iv) compression (ubc) for images (a) 1-2 and (b) 1-4 from the dataset. The image degradation is much higher in 1-4 pair than in the 1-2 pair. Columns 1 and 2 are results on the Harris-affine and Hessian-affine detectors respectively and the EMD matcher of [21] is used for all results in this figure. Note that the scales are different for different figures to improve the clarity of the plots.



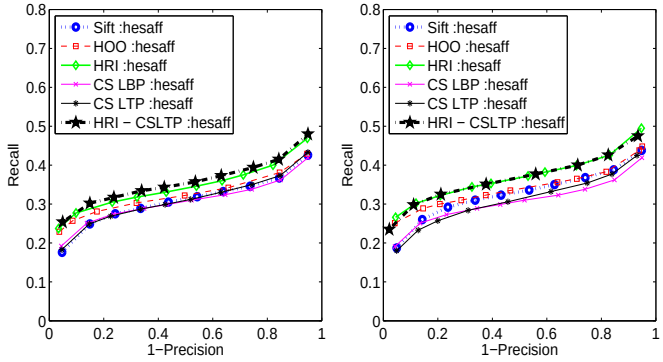
(affine change)



(illumination change)



(blur)



(a) (b)

(compression)

Figure 5. Comparison of different methods on different degradations for images 1-2 for Hessian-affine detectors for (a) L_2 and (b) EMD matcher of [21].

and photometric transformations and for different scene types. Six different transformations are evaluated: viewpoint change, scale change, image rotation, image blur, illumination change, and JPEG compression.

The evaluation criterion is based on the number of correct and false matches between a pair of images. The number of correct matches is determined with the *"overlap error"* [13]. A match is assumed to be correct if the ovelap area is > 0.5 of the union of the two areas of the descriptors. We use this criteria for all results in this section.

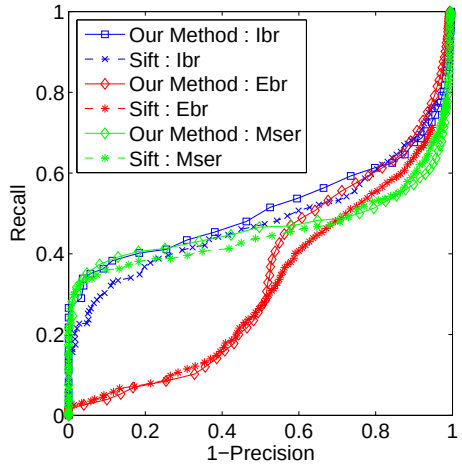
The results of our combined method when compared with existing techniques are shown in Figure 4. For these plots, the matching method used is the EMD matcher from [21] as it gave better results than the L_2 distance for all methods with not so much increase in the running time (about double). We performed better than almost all existing methods for almost all cases (7 out of 8 cases, although only 4 are shown here). For illumination change (leuven), the problem seems to be inconsistent normalization as our method assumes a linear change in the intensities and the very low intensities encountered in this set of images leads to non-linearity close to the lower range of the sensor. The adaptive normalization technique helped improve the results for this case, although still the results were still not consistent with results on other images.

In Figure 5, we show some more details of our method by showing the results separately for the two descriptors and how the combination of the two gives results superior to both. We also show the results using the L_2 measure which may be compared to the results obtained using the EMD matcher which are better for all the methods. We have also compared our results with HRI + SIFT which performs lower then our current results.

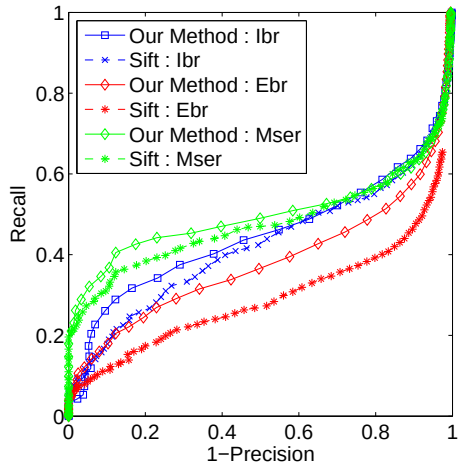
Finally, in Figure 6, we show the result of our matcher on some other detectors: MSER, IBR and EBR for the leuven (illumination change) and bikes (blur) image sets. As can be seen, the relative performance of the different methods on these detectors is similar to Hessian-affine and Harris-affine. Similar behavior was found for other image sets as well.

6.2. Caltech 3D dataset

The second dataset that we tested our algorithm on is the 3D dataset from Caltech[17]. The tests on this dataset mimic the Object Recognition problem and should give us some idea of the performance of our detector for this popular and important problem. Objects are put on a turntable and keypoints are matched for different rotation angles. They are also matched with points from a random database. If the distance of the keypoint from its best matching keypoint is less than a factor of its distance to the second best match, then it is accepted as a matched point. Then, it is tested whether the matched keypoint comes from an image



(a) leuven



(b) bikes

Figure 6. Comparison of our matching approach with SIFT for illumination and blur changes for image pair 1 and 2 from the Oxford dataset using different detectors.

of the same object and satisfies the epipolar constraint(s). It is taken as a correct match if it satisfies both these criteria, else it is flagged as a false match.

Since Hessian-affine combined with SIFT was shown to give the best results in [17], we have shown the comparative results only on Hessian-affine and against SIFT. These results are shown in Fig 7 and Fig. 8 for the L_2 and EMD[21] metrics respectively. The detection rate is plotted against the false alarm rate. The detection rate is the number of correct matches divided by the total number of matches tried while the false alarm rate is the ratio of wrong matches to the total number of matches. Also shown in the figures is the detection rate as a function of the viewing angle for a false alarm rate of 0.01. As was observed for the Oxford dataset, we got substantial improvement when the false alarm rate was low but the results are close to SIFT at the higher false alarm rates. Again, since mostly we would like to work at low false alarm rates, the improvement in performance in significant. More information on this dataset and these plots can be obtained from [17] as we have followed their convention for the results.

7. Conclusions and Future Work

We have presented two different order-based methods for feature description: the Histogram of Relative Intensities (HRI) and Center-Symmetric Local Ternary Patterns (CS-LTP). These methods were designed to be more robust to Gaussian noise than previous methods based on orders. This was achieved by considering intensity information along with order information rather than using only order information as in previous methods. While the individual results of the two methods developed were encouraging in themselves, the combination of the two gave results better than either of them individually due to the orthogonal nature of the two descriptors. Better performance than gradient-based approaches is perhaps due to the more stable nature of the raw intensities compared to raw gradients, especially in the presence of image degradations such as affine transformation, image blur and image compression. The EMD distance measure of [21] improves the matching accuracy for all methods and we recommend it as the match measure instead of the commonly used L_2 distance.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *CVIU*, 110(3):346–359, June 2008.

[2] D. Bhat and S. Nayar. Ordinal measures for image correspondence. *PAMI*, 20(4):415–423, Apr. 1998.

[3] N. C. Feng T., Suk Hwan L. and H. T. A novel feature descriptor invariant to complex brightness changes. *CVPR*, 2009.

[4] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988.

[5] M. Heikkila, M. Pietikainen, and C. Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425–436, Mar. 2009.

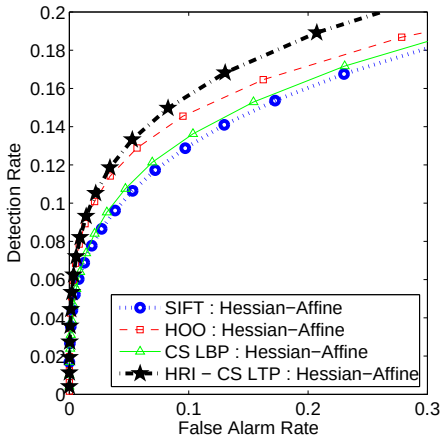
[6] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *ECCV*, pages Vol I: 228–241, 2004.

[7] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *CVPR*, pages II: 506–513, 2004.

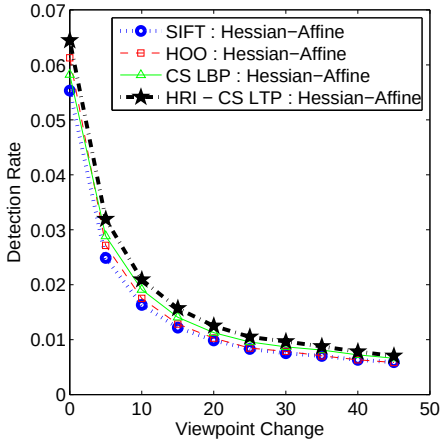
[8] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *PAMI*, 27(8):1265–1278, Aug. 2005.

[9] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *PAMI*, 29(5):840–853, May 2007.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov. 2004.

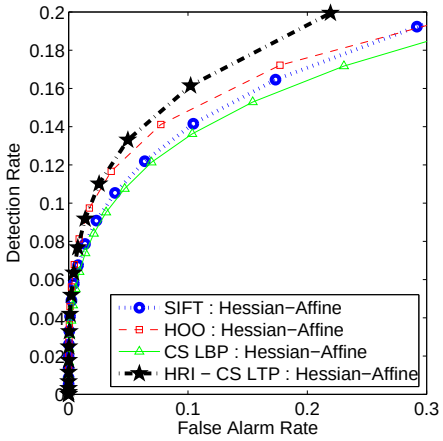


(a)

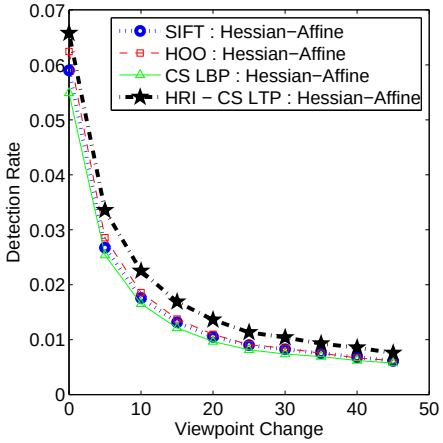


(b)

Figure 7. Results of experiments on the 3D dataset for Hessian-affine with L_2 metric. (a) Detection Rate vs. False Alarm rate, (b) Detection Rate vs. Rotation angle at false alarm rate of 0.01.



(a)



(b)

Figure 8. Results of experiments on the 3D dataset for Hessian-affine with EMD metric. (a) Detection Rate vs. False Alarm rate, (b) Detection Rate vs. Rotation angle at false alarm rate of 0.01.

[11] T. Matthew and W. W. Sift-rank: Ordinal description for invariant feature correspondence. *CVPR*, 2009.

[12] K. Mikolajczyk and J. Matas. Improving descriptors for fast tree matching by optimal linear projection. In *ICCV*, pages 1–8, 2007.

[13] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, page I: 128 ff., 2002.

[14] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, Oct. 2004.

[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, Oct. 2005.

[16] A. Mittal and V. Ramesh. An intensity-augmented ordinal measure for visual correspondence. In *CVPR*, pages I: 849–856, 2006.

[17] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. *IJCV*, 73(3):263–284, July 2007.

[18] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *PAMI*, 27(11):1832–1837, Nov. 2005.

[19] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, July 2002.

[20] T. Pajdla, M. Urban, O. Chum, and J. Matas. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, page 3D and Video, 2002.

[21] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *ECCV*, pages III: 495–508, 2008.

[22] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, Nov. 2000.

[23] M. Singh, V. Parameswaran, and V. Ramesh. Order consistent change detection via fast statistical significance testing. In *CVPR*, pages 1–8, 2008.

[24] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *Analysis and Modelling of Faces and Gestures*, pages 168–182, 2007.

[25] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *IJCV*, 59(1):61–85, Aug. 2004.

[26] R. Zabih and J. Woodfill. Non-parametric local transforms fo computing visual correspondence. In *ECCV*, pages 151–158, 1994.