2015 IEEE International Conference on Robotics and Automation (ICRA)
Washington State Convention Center
Seattle, Washington, May 26-30, 2015

LOIND: An Illumination and Scale Invariant RGB-D Descriptor*

Guanghua Feng[1], Yong Liu[2], Yiyi Liao[1]

Abstract— **We introduce a novel RGB-D descriptor called local ordinal intensity and normal descriptor (LOIND) with the integration of texture information in RGB image and geometric information in depth image. We implement the descriptor with a 3-D histogram supported by orders of intensities and angles between normal vectors, in addition with the spatial sub-divisions. The former ordering information which is invariant under the transformation of illumination, scale and rotation provides the robustness of our descriptor, while the latter spatial distribution provides higher information capacity so that the discriminative performance is promoted. Comparable experiments with the state-of-art descriptors, e.g. SIFT, SURF, CSHOT and BRAND, show the effectiveness of our LOIND to the complex illumination changes and scale transformation. We also provide a new method to estimate the dominant orientation with only the geometric information, which can ensure the rotation invariance under extremely poor illumination.**

I. INTRODUCTION

In the fields of computer vision and robotics, feature matching is the essential operation in many complex problems, such as image stitching, wide baseline matching, coarse alignment and image retrieval etc. The basic idea of image based feature matching is to find some local interest points or regions in images by detectors, and then encode those points or regions by relative invariant descriptors for further matching. Excellent descriptor should be robust to illumination, scale, noise and other complex environments.

As the RGB images are rich in texture information and have high resolutions, researchers have presented many discriminative descriptors, such as SIFT [1] and SURF [2], which are widely applicable and robust to many variations and distortions. However, under some extreme cases such as dramatic changes of lighting and textureless scenes, those classical descriptors cannot always achieve enough discriminative performance. Meanwhile, 3D surfaces can be built by laser or other high-precision TOF sensors. There are also some descriptors generated from those dense depth point data, such as Spin Images [3], which is a prominent histogram descriptor based on geometric encoding and is invariant to rotation. SHOT descriptor [4], which designs robust and unambiguous local RF (Reference Frames) to encode descriptors. The drawbacks of depth based 3D descriptor are obviously, they require high quality data points to maintain their discriminatively capability. Once the scene only contains flat surfaces, their performances will decrease dramatically. Moreover, their computational complexities are also high.

Nowadays, there are some new works focusing on combining both RGB and Depth information to construct uniform descriptors, e.g. CSHOT [16] and BRAND [11]. Although it has been proved that combination may provide better performance than any single of one's [15], current RGB-D combination based descriptors also have some limitations. CSHOT's performance is highly dependent on the accuracy of the geometric information, thus it cannot achieve well performance on low quality depth data, such as data from Kinect. BRAND may suffer from the complex changes between two images due to its lack of texture information.

In this paper, we combine both texture and low accurate geometric information, and propose a local ordinal intensity and normal descriptor (LOIND). The essential idea is to take both advantages of the RGB information and the Depth information. On the one hand, the depth data is utilized to serve as the main discriminative metrics when the scenes have image blur, weak illumination and less texture information. On the other hand, we utilize rich texture information to reduce errors introduced by rough depth information.

The following Section II provides an overview of the related works on local invariant descriptors. Our LOIND descriptor is explicitly presented in Sections III. Finally, experiments are described in Section IV followed by concluding remarks and future works in Section V.

II. RELATED WORKS

For RGB descriptors, there are two main categories of descriptors, which are based on the relative values and absolute values respectively. Relative methods try to construct descriptors according to the point-pair-wise comparison of the intensity. For example, Brisk[5] is composed as binary strings with the results of intensity comparison between randomly chosen point pairs. As for methods based on absolute values, they usually construct histograms based on intensities or gradient orientations. For instance, descriptors such as SIFT [1] and GLOH[6] compute the histograms of gradient orientations and locations, while SURF [2] computes the histogram of Haar wavelet. Generally speaking, the histogram-based methods perform better than the binary-based or moment-based descriptors because of the richer information that the histogram contains. Several other descriptors are proposed to combine these two ideas such as OSID [7] and LIOP [8]. They compute histograms to combine the relative ordering of the pixel intensities and spatial information, which are verified to improve upon both of them.

*Research supported by National Natural Science Foundation of China (61173123)Zhejiang Provincial Natural Science Foundation of China (LR13F030003).

[1]G.H. Feng and Y.Y. Liao are with the Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang,310027, China.

[2]Y. Liu is with State Key Lab of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, 310027, China (He is the corresponding author of this paper, e-mail: yongliu@iipc.zju.edu.cn).

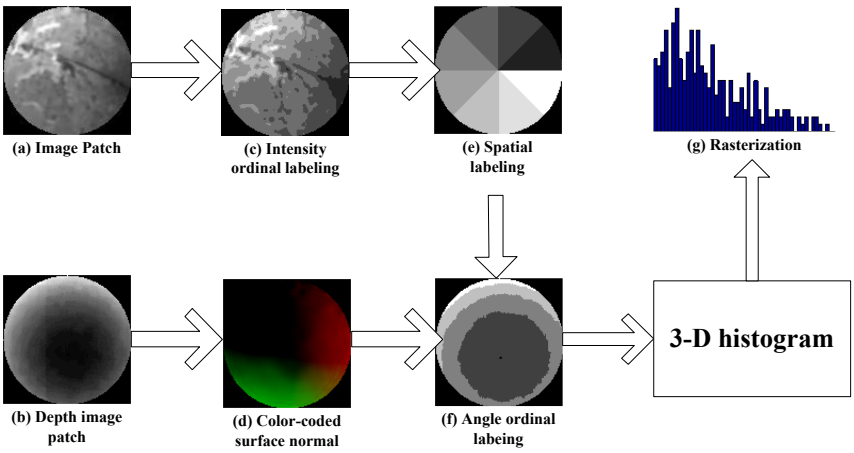


Figure 1. The general diagram of our LOIND descriptor

However, those 2D descriptors have the inherent limitations, such as the requirement on rich texture information and stable lighting condition.

With the depth information, we can construct descriptors under the insufficient texture information and severe lighting conditions. The depth descriptor can be divided into two categories[9], i.e. histogram descriptors and signature descriptors. The former ones construct a histogram with respect to the neighboring points, while the latter ones calculate some values individually of the neighboring points and encoders. For example, Histogram of Normal Orientations[10] is a normal histogram storing the angle between the reference axis and the normal of a neighboring point. Besides, SHOT(Signature of Histograms of Orientations)[4] is proposed to combine the histograms and signatures for better performance and stability. The depth descriptors are usually more computationally expensive and more reliable to the precision of hardware.

In recent years, the descriptors combining varied information sources are carried out to enhance the adaptability under different image transformation. CSHOT combines SHOT descriptor with texture-based descriptor by the way of combining histogram and signature. There is also a descriptor cascades Spin Image and SIFT [1], it has been validated the effectiveness of this combined descriptor will perform better than either texture information or depth information alone. BRAND is another efficient binary descriptor which combines BRIEF descriptor[12] and depth information.

III. Our Method

Our descriptor is based on the idea of relative order information in both the appearance and depth information. The descriptor is constructed by a three-dimensional space, shown in figure 1, i.e. intensity, the angle between normal vector and the spatial structure, corresponding to texture, depth and patch's spatial structure information respectively.

A. Pre-processing and Feature Detector

Since the process of ordinal intensity and ordinal angel is in the pixel level, they are noise-sensitive. So we need to smooth the patches and increase stability and repeatability of descriptors. Both the RGB and depth images are smoothed by the Gaussian filter. A sour descriptor is encoded by the intensity ordering, the local intensity-extreme detector, such as Harris-affine or Hessian-affine, is used to detect the interest point from the RGB images. The detected interest point is also employed to the depth image by the coordinate transforming.

B.Scale estimation

According to the principle of image formation, the scale of a tiny region is approximately inversely proportional to the corresponding depth. Thus we can estimate the scale of each feature point based on the geometrical information using a scale factor *s*. For the local patch, with a size of $r \times r$, chosen based on the point's depth value, its scale is computed by:

$$s = \max\left(0.2, \frac{3.4 - 0.4\max(1,d)}{3}\right)$$

where d is the depth value of keypoint. When $d \leq 1\text{m}$ or $d \geq 7m$, we need to limit the value of r since Kinect's measurement is not accurate in this range. By the scale value we can infer the size of patch:

$$r = R * s$$

R is the maximum threshold of the radius. It is an empirical value according to our experiment, when R=[20,70], image shows the best effect. If scale varies gently, we can choose a smaller R. The value of R is roughly valuated as:

$$R = -5 + 25 * \min\left(3, \frac{\max(0.2, s_{max})}{\max(0.2, s_{min})}\right)$$

s_{max} and s_{min} are the maximum and the minimum scale values in the image. The number of pixels in our patch is R^2 .To reduce the computational complexity, we normalize each patch by shrinking all the patches to a region of fixed size, the value of normalization radius is set to*20* in our experiment.

C. Ordinal and Spatial Labeling

To generate a discriminative RGB-D descriptor, we further divided the process into two parts, the encoding of spatial distribution and the method of encoding RGB-D information. Firstly, the encoding of spatial distribution aims to divide the certain range of neighborhood of interest points into several sub-regions, under the assumption of the relatively local invariant property at the key point. It increases the dimension of the descriptor and the information entropy attached in each interest point, so the patch can be

discriminated by these featured descriptors. Afterwards, the RGB-D information in each patch of the interest point is encoded into descriptor, and the method of the encoding process determines the characteristics of the descriptor and its applicable scene.

Our LOIND divides the local patch into several equal sub-regions in the RGB image, the sub-regions division in depth image is the same as the RGB image. The encoding is then established from the relative ordering space by building the histogram from three dimensions, i.e. the spatial space, relative ordering space of the intensity and relative ordering angle space of the normal vectors.

Spatial distribution

We first denote the interest point as the center and extract a circle patch around it with the radius of *d* in RGB image plane. Then we divide the patch into *npies* sectors, and calculate the pixel distribution in each sector. Each sector can be regarded as an encoding unit in the spatial distribution, we then encode the intensity and depth information for each sector in the relative ordering spaces.

Encoding for ordinal distribution

After dividing the patch around the interest point, we can encode the information in each sector. All the encoding is adapted by pixel's position, intensity and depth values based on relative ordering.

Instead of building the statistical histogram in the absolute intensity space, our method transfers the intensity values to the relative intensity space with respect to the center interest point for each pixel in the patch, shown in figure1(c). All the pixels in the sector are then sorted based on their relative intensity values with respect to the center points. According to the ranking of all the pixels in the patch, we group the relative intensity values into *nbins* bins, where each bin has the same number of the relative intensity values. For example, if there are *N* different relative intensity values in the relative intensity space which is divided into *nbins* bins, then each bin contains *N/nbins* relative intensity values. The histogram can be generated by calculating the number of the pixels whose relative intensity values are belonging to the corresponding bins. This relative ordering based histogram can avoid inconstant intensity values caused by illumination changes. Although the absolute intensity value changes due to varied illuminations, the ranking value is usually constant.

The intensity sorting is implemented by quicksort. In order to reduce the computational complexity, the selection algorithm is used and thus it can directly obtain the boundary value of each bin.

Ordinal angle between normal vectors

We first calculate the normal vector of each pixel in the patch to construct the surface normal space, shown in figure 1(d), and then calculate the dot product between normal vector of the center points and the normal vectors of other points in the patch, as shown in figure2.

The dot product of two vectors is calculated as:

$$\mathrm{d}(x_i) =< p_n(x_i), p_n(x_o) > \quad (1)$$

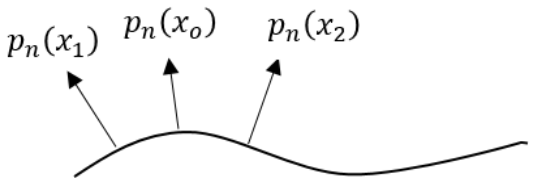


Figure 2. Normal vector in depth space

$p_n(x)$ is the normal vector of point x, x_o represents the center of the patch. θ_i is the angle between $p_n(x_i)$ and $p_n(x_o)$, then dot product value $d(x_i)$ is inversely proportional to θ_i.When θ_i=0,$d(x_i) = 1; \theta_i = \pi$, $d(x_i) = -1$.When $\theta_i \approx 0$, considering the curve cross two points can be nearly treated as a plane, the result may be affected by rough depth value. Thus we set:

$$\mathrm{d}(x_i)>=\rho, dbin(x_i)=1; \quad (2)$$

where ρ is a threshold to determine whether the two-point concave surface approximate flat. For all the dot product values that $\mathrm{d}(x_i)<\rho$,we will rank the values of the dot product with respect to the center point, then the sorted sequence is divided into *dnbin* bins based on the values. Thus the depth distribution can be divided into *dnbin*+1 bins, *dnbin* bins for the sorted values and one additional bin for all those plane pixels ($\mathrm{d}(x_i)>=\rho$). This method encode the depth information into descriptor can greatly improve the descriptor's discriminative capacity, and perform better in the circumstance like textureless scene.

D. Descriptor Construction

In our LOIND, each local patch is constructed to a histogram from three dimensions, where *X* axis represents intensity relative ordering, *Y* axis is the spatial distribution and *Z* axis is the dot product based relative ordering, i.e. the relative ordering of angles. Then we transform the three dimension histogram into a vector, and the total dimension is *nbins***npies**(1+*dbins*).

The combination sequence of the intensity, spatial and angle will not influence the discriminative capability. Generally the ordering variable is arbitrary. However, inconsistent combination sequence may increase the probability of miss-matching. So we define the combination sequence of x, y, z as intensity, spatial and angle. The relative intensity values and relative angles between normal vectors are sorted from small values to large values, the sectors (pies) of the circled patch is sorted anticlockwise. In order to eliminate the effect caused by different patch size, we also normalize the descriptor vector into the values of ratio.

IV. EXPERIMENTS

In this section, the comparable experiments on stat-of-art descriptors are carried out to validate the effectiveness of the proposed descriptor.

A. Dataset and Evaluation Metric

In the experiments, we use the public RGB-D datasets[1,2] as our benchmark data. For clarity, we only show the comparison results on two image pairs chosen from these two datasets

[1] https://cvpr.in.tum.de/data/datasets/rgbd-dataset
[2] http://rgbd-dataset.cs.washington.edu/

(a) original image (b) image with linear change

Figure 3. Image pairs with natural illumination changes

(a) image with square root change (b) image with square change

Figure 4. Image pairs with synthesized illumination changes

respectively, which are the handle slam sequence *freiburg2_desk* and the RGB-D Scenes sequence *kitchen_small_1*.

Except for the publicly available images, we also capture image pairs of "dog" in different linear lighting conditions as shown in figure 3. To further evaluate the performance under complicated lighting conditions, we synthesize two images by performing square root and square illumination change with respect to figure 4. These nonlinear transformations bring huge challenges to the descriptor.

We employ four state-of-art descriptors in the comparison experiments. As for evaluation, we use the metric proposed by Mikolajczyk and Schmid [6]. We perform the threshold based matching on all pairs of keypoints in the corresponding image pairs. If the Euclidean (for SIFT, SURF, CSHOT) or Hamming (for BRAND) distance to the nearest neighbor is below a threshold *t*, this pair is regarded as valid match.

B. Parameter setting

We discuss the influences of our parameters by experiments to find the appropriate parameter setting.

The sigma in smoothing: Experimentally, we found that image smoothing can significantly reduce noise and improve the performance of our descriptors. As validated in the experiments, we use 5*5 Gaussian kernels and set sigma=1 in the 2D images. As for the noisy depth image, we use 10*10 kernels with sigma=2.

Normal vector estimation: Normal vector can be estimated by PCA on the nearest neighbors [13].For each point, we centralize the 3D coordinates of its nearest neighbors, and calculate the covariance matrix and the corresponding eigenvalues. Then the normal vector is the eigenvector with respect to the minimum eigenvalue. However, this method is time-consuming. Since the data of Kinect can be transformed to organized point cloud, we employ the integral image method [14], which is more efficient and accurate. Figure 5 shows that the performance of

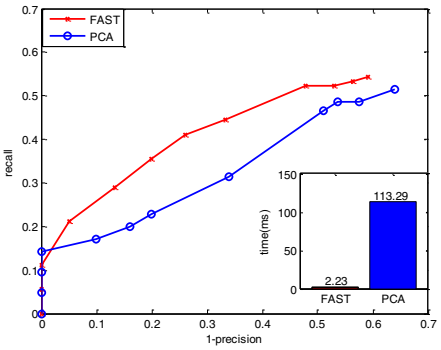


Figure 5. Precision-recall performances under different normal vector estimation methods.

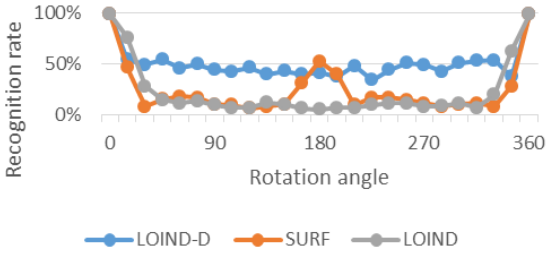


Figure 10. Recognition rate when matching the image against a rotated image at every 15 degrees.

the integral image method is better than that of PCA. Moreover, it has significant superiority on the computational efficiency.

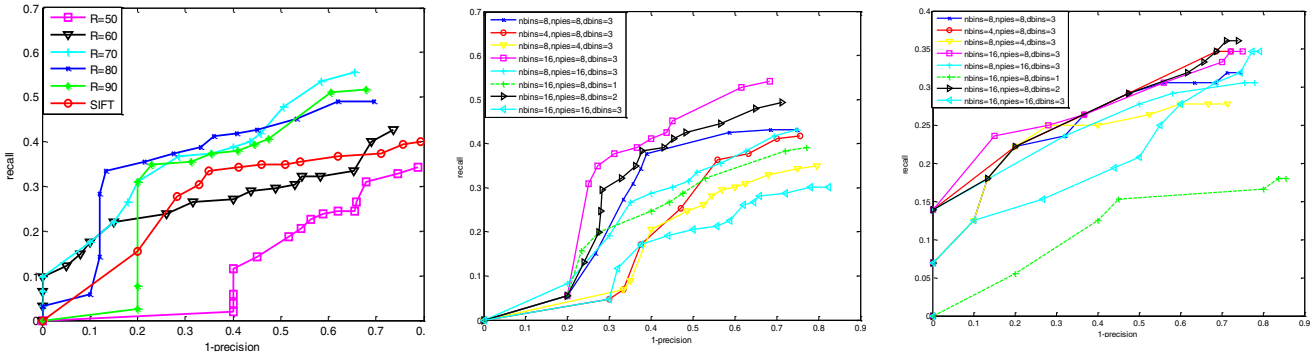
Scale factor: The scale factor is chosen according to the depth information, the radius of the patch is calculated with *r=R*s*. We use an image pair from *freiburg2_desk*, there is a triple scale transformation between these two images. The performance with different *R* is shown in figure 6(a). Considering both matching performance and calculation efficiency, we choose *R=70* as the best parameter value. If the scale change of the image pair is slighter, we can use smaller R for more efficiency.

Numbers of ordinal, spatial and depth bins: We choose two image pairs to validate the influence of those parameters. Three parameters are set as $nbin \in (4,8,16)$, $npies \in (4,8,16)$, and $ndbin \in (1,2,3)$ respectively. When *nbin*=1, it equals to ignore the affection of depth information. As can be seen, the performance is promoted by adding the depth information. More patches we use, the descriptor dimension will be higher, which is able to contain more texture and depth information. However, too many patches will lead to high complexity in matching, and sensitive to noises.

Figure 6(b) and 6(c) show the performances of varied combinations of the *nbin*, *npies* and *ndbins*. The performance of *nbins* = 16 is better than that of *nbins*=4, 8, while *npies*=8 performs bestand *ndbins* performs better when setting as 3 or 2. To balance the efficiency and performance, we choose the dimension as (*nbin*npies*ndbins*)= 8*8*3=192.

C. Performance evaluation

We compare the matching performance of our LOIND and the other four methods as shown in figure 7, 8, 9. For all the descriptors, keypoints or regions are detected by harris-affine



(a) (b) (c)

Figure 6. Precision-recall performance under different parameters. (a) Performances under different normal vector estimation methods. (b)(c) Performances under different scale factors, where (b) is performed on freiburg2_desks sequence and (c) is performed on kitchen_small_1sequence

method. We have selected the same two sequences from the datasets which include scale transform, viewpoint change and noise depth and texture. The precision-recall curves in figure 7 and figure 8 show that our LOIND outperforms the other methods on each datasets, and figure 9 shows that our LOIND also performs better on the image pairs we captured above with the non-linear illumination changes.

D. Orientation estimation

To realize rotation invariance, we normally use the RGB texture information to calculate the dominant orientation. If the lighting condition is poor and the texture information is missing, we propose to estimate the dominant orientation completely according to depth information. In our approach, we find the pixels with $d(x_i)<\rho$ and calculate the dot product between the normal vectors attaching with these pixels and the normal vector of the central point for each pie. Then we obtain the pie with smallest mean dot product, which means this pie has the largest angle to the central point. The dominant orientation is then denoted as the orientation of this pie. The starting pie in spatial distribution should be set from the patch's dominant orientation. In figure 10, we test the recognition rate of rotation under poor lighting and noisy texture situation, results show that our method performs well and stable.

V. Conclusion and Future Works

LOIND is a novel RGB-D descriptor that is invariant to scale transform and robust to linear or nonlinear brightness changes. The core idea is relative ordering, encoding both texture and depth information. In fact, most of the transform will not affect the intensity ordering or geometric structure. Thus experiments shows that LOIND can achieve better performance than state-of-art descriptors, including typical RGB descriptors, SIFT, SURF, and the RGB-D descriptors e.g. BRAND and CSHOT, which are also based on the combination of geometric and appearance. Meanwhile, when the situation that illumination is extremely weak, we propose a dominant direction estimation method from purely depth information, our method can provide accurate dominant direction when the illumination is extremely poor or there is less texture information.

The main limitation of RGB-D descriptor lies on the accuracy of depth sensor. So the future work may focus on encoding discriminative RGB-D descriptor without losing that information and provide adaptive RGB-D descriptor driving by data.

References

[1] David G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, November 2004.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. ECCV, 2006.

[3] A. E. Johnson and M. Hebert. Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes,PAMI, 433–449, 1999.

[4] F. Tombari, S. Salti, L. Di Stefano, "Unique Signatures of Histograms for Local Surface Description", ECCV, September 5-11, 2010.

[5] Leutenegger,S.Chli, M. ; Siegwart, R.Y.BRISK: Binary Robust invariant scalable keypoints,ICCV,Nov.2011,2548-2555.

[6] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors. In PAMI 27(10):1615-1630 , 2004.

[7] Tang F, Lim S H, Chang N L, et al. A novel feature descriptor invariant to complex brightness changes,CVPR 2009, 2631-2638.

[8] Zhenhua Wang , Fan, Bin , Fuchao Wu . Local Intensity Order Pattern for feature description.ICCV,Nov 2011:603-610.

[9] Jens Behley, Volker Steinhage and Armin B. Cremers,Performance of Histogram Descriptorsfor the Classification of 3D Laser Range Datain Urban Environments,ICRA,May 2012,4391-4398

[10] R. Triebel, K. Kersting, and W. Burgard, "Robust 3D Scan Point Classification using Associative Markov Networks," in ICRA, 2006,pp. 2603–2608.

[11] Nascimento,E.R.Oliveira, G.L. ; Campos, M.F.M. ; Vieira, A.W. ; Schwartz, W.R. BRAND: A robust appearance and depth descriptor for RGB-D images .IROS,Oct.2012,1720-1726

[12] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary RobustIndependent Elementary Features," in ECCV, September 2010.

[13] J. Berkmann and T. Caelli, "Computation of surface geometry and segmentation using covariance techniques," IEEE Trans. PAMI, vol. 16,no. 11, pp. 1114–1116, nov 1994.

[14] S. Holzer, R. B. Rusu,M. Dixon ,S. Gedikli,N. Navab.Adaptive Neighborhood Selection for Real-Time Surface NormalEstimation from Organized Point Cloud Data Using Integral Images. IEEE/RSJ International Conference on Intelligent Robots and Systems,Oct.2012.2684-2689.

[15] K. Lai, L. Bo, X. Ren, and D. Fox, Sparse distance learning for objectrecognition combining rgb and depth information, in ICRA, 2011

[16] F. Tombari, S. Salti, L. Di Stefano, A combined texture-shape descriptor for enhanced 3D feature matching, IEEE International Conference on Image Processing (ICIP), September 11-14, Brussels, Belgium, 2011

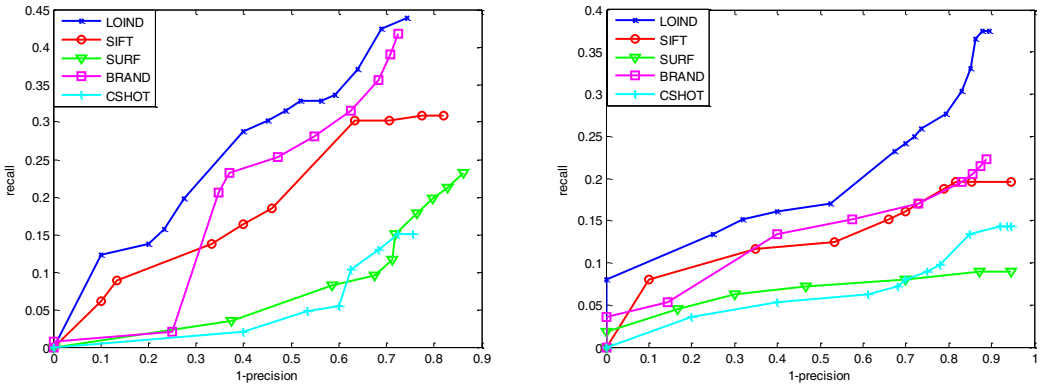


Figure 7. Comparison results on *freiburg2_desks* data set with two image pairs

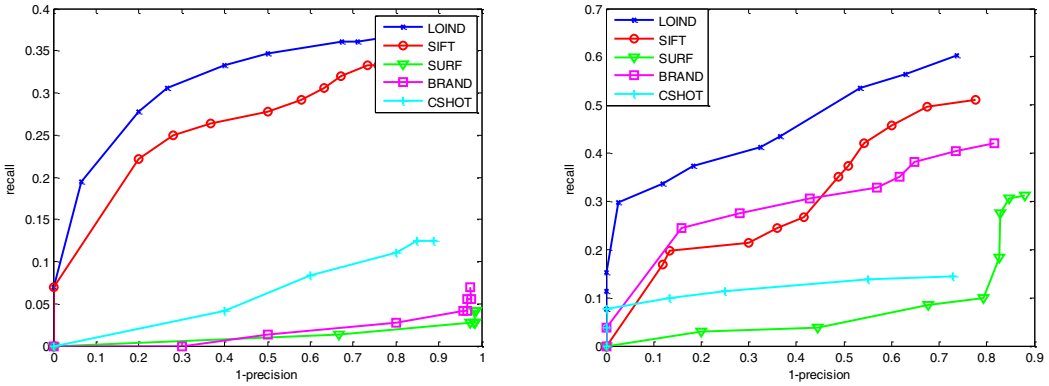
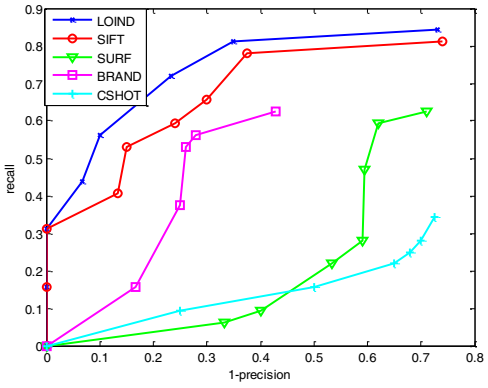
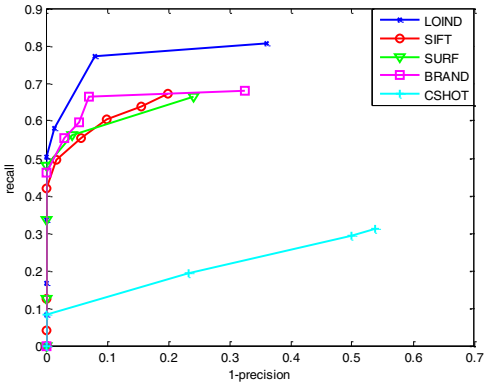


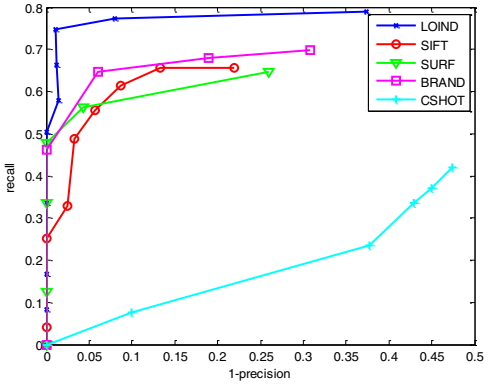
Figure 8. Comparison results on *kitchen_small_1*data set with two image pairs



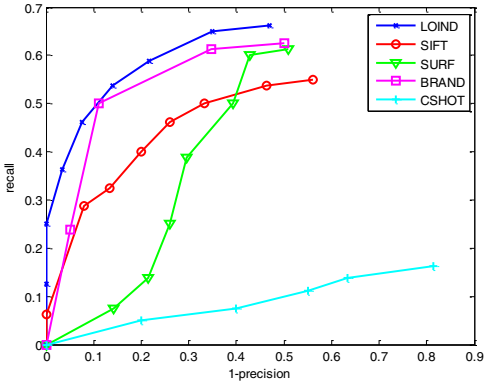
(a) Image pair of figure 3(a) and (b)



(b) Image pair of figure 3(a) and figure 4 (a)



(c) Image pair of figure 3(a) and figure 4(b)



(b) Image pair of figure 4(a) and figure 4 (a)

Figure 9. Comparison results on varied illumination conditions