2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

BRAND: A Robust Appearance and Depth Descriptor for RGB-D Images

Erickson R. Nascimento Gabriel L. Oliveira Mario F. M. Campos
Antônio W. Vieira William Robson Schwartz

Abstract— **This work introduces a novel descriptor called Binary Robust Appearance and Normals Descriptor (BRAND), that efficiently combines appearance and geometric shape information from RGB-D images, and is largely invariant to rotation and scale transform. The proposed approach encodes point information as a binary string providing a descriptor that is suitable for applications that demand speed performance and low memory consumption. Results of several experiments demonstrate that as far as precision and robustness are concerned, BRAND achieves improved results when compared to state of the art descriptors based on texture, geometry and combination of both information. We also demonstrate that our descriptor is robust and provides reliable results in a registration task even when a sparsely textured and poorly illuminated scene is used.**

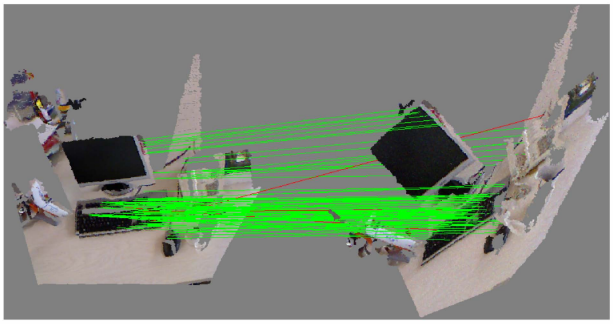


Fig. 1. Three-dimensional matching example for two scenes using BRAND descriptor. Mismatches are shown with red lines and correct matches with green lines.

I. Introduction

At the heart of numerous tasks, both in robotics and computer vision, resides the crucial problem known as correspondence. Simply stated, the key challenge is to automatically determine for a given point in one image, a point in another image which is the projection of the same point in the scene. This is a challenging problem due to several issues, such as scene illumination, surface reflectance, occlusion and acquisition noise.

Developing accurate three-dimensional models of scenes, Simultaneous Localization And Mapping (SLAM), tracking, and object recognition are representative examples of applications that build upon a correspondence foundational layer, of which a good descriptor is the cornerstone.

In recent years, two-dimensional images have been used since they provide rich textural information, which allows the development of several feature descriptor approaches. Scale Invariant Feature Descriptor (SIFT) [1] and Speed Up Robust Descriptor (SURF) [2] are the most popular and representative methods of these approaches. Nonetheless, common issues concerning real scenes, such as variation in illumination and textureless objects, may dramatically decrease the performance of these descriptors.

With the growing availability of inexpensive, real time depth sensors, depth images are becoming increasingly popular. As in the case of two-dimensional images, region matching on this type of data is most advantageous. Spin Images [3] is a classical example of a robust and rotation invariant geometric descriptor, successfully employed in applications such as object recognition [4]. However, due to the geometric nature of the data, such descriptors tend to present high complexity and large ambiguous regions may become a hinderance to the correspondence process.

The combination of appearance (provided by two-dimensional textured images), and geometric (produced by depth information) cues, has proven to be a very promising approach to improve the correspondence and recognition rates. Lai et al. [4] have already shown that the concatenation of two well-known descriptors for each type of data (SIFT and spin images) to form a new descriptor, outperforms learning from view-based distance using either appearance or depth alone.

In this paper, we present a novel local RGB-D descriptor called BRAND, which efficiently combines intensity and geometric information to substantially improve discriminative power enabling enhanced and faster matching. Different from descriptors that use either appearance information or geometric information, our approach proposes to build a single descriptor which simultaneously takes into account both sources of information to create an unique description of a region. Figure 1 shows the correspondence of a set of keypoints in two 3D scenes achieved using BRAND. Experimental results show that BRAND is a robust and computationally efficient technique that outperforms state-of-the-art techniques in accuracy, processing time and memory consumption.

The authors are affiliated with the Computer Vision and Robotic Laboratory (VeRLab), Computer Science Department, Universidade Federal de Minas Gerais, MG, Brazil. This work has been supported by grants from CNPq, CAPES and FAPEMIG. E-mails: {erickson,gabriel,mario,william}@dcc.ufmg.br
Antônio Wilson Vieira is also affiliated to CCET, Unimontes, MG, Brazil. E-mail: awilson@dcc.ufmg.br

II. Related Work

Computer vision literature presents numerous works on using different cues for correspondence based on textural information. SIFT [1] and SURF [2] are popular algorithms for keypoint detection and descriptor creation. They build their feature detectors and descriptors on local gradients and specific directions to achieve rotational invariance. More recently, several compact descriptors, such as [5], [6], [7], [8] have been proposed employing ideas similar to those used by Local Binary Patterns (LBP) [9]. These descriptors are computed using simple intensity difference tests, which have small memory consumption and modest processing time in creation and matching process. However, in virtually all of these approaches, features are extracted from images alone, and as consequence, they are more sensitive to variations in illumination and they are not able to handle images of textureless objects.

If on one hand image texture information can usually provide better perception of object features, on the other hand depth information captured by 3D sensors is less sensitive to lighting conditions. Spin images descriptor [3] is an example of a successful descriptor extracted from 3D data. This approach creates a rotation invariant 2D representation of the surface patch hemming a 3D point. Other approaches proposed to handle unordered 3D point clouds are based on feature histograms [10], [11]. Even though these descriptors are accurate, they present high computational cost since the construction of a single descriptor for general raw point clouds or range images involves highly complex geometric operations.

A promising idea for the design of descriptors that is becoming popular in the last few years is to consider multiple cues. Zaharescu et al. [12] proposed the MeshHOG descriptor using texture information of 3D models as scalar functions defined over a 2D manifolds. Tombari et al. [13] presented the Color-SHOT (CSHOT) descriptor based on an extension of their shape only descriptor Signature of Histograms of Orientations (SHOT) [14] to incorporate texture. CSHOT signatures are composed of two histograms, one contains the geometric features over the spherical support around the keypoint and the other contains the sum of the absolute differences between the RGB triples of the each of its neighboring points. The authors of [13] compared CSHOT against MeshHOG and reported that their approach outperformed in processing time and accuracy. In the case of global descriptor, Kanezaki et al. [15] presented the Voxelized Shape and Color Histograms (VOSCH) descriptor, which by combining depth and texture, was able to increase the recognition rate in cluttered scenes with obstruction.

Considering the promising ideas employed in [15], [13], [12], our proposed local descriptor brings forth the advantages of using both appearance and depth information. However, differently from them, our approach spend little memory space and little processing time without losing accuracy.

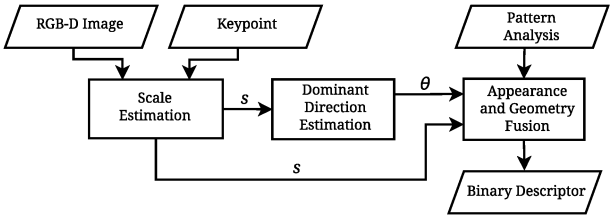


Fig. 2. Assembling diagram of BRAND descriptor. After computing the scale factor s using depth information from RGB-D image, our methodology extracts a patch of the image in the RGB domain to estimate the dominant direction θ of the keypoint. Finally, appearance and geometric information are fused based on the features selected with a pattern analysis.

III. Methodology

In this section we detail the design of the BRAND descriptor. The stages performed to build this descriptor are illustrated in Figure 2 and will be described in detail in this section.

Our methodology, which is composed of three main steps, receives a list of keypoints that can be detected by algorithms such as [1], [2], [16], [17], and returns a list of signature vectors. In the first step, we compute the scale factor using the depth information from RGB-D image. The scale factor is used in the next step (dominant direction estimation) and in the feature analysis in the keypoint's vicinity. In the dominant direction estimating step, a patch in the RGB domain is extracted and used to estimate the characteristic angular direction of the keypoint's neighborhood. At last, we combine both appearance and geometric information to create keypoints descriptors that are robust, fast and lightweight. The goal is to bring forth the best cues that each domain can provide, and combine them as efficiently and as inexpensively as possible, into one binary string.

A. Scale and Orientation Assignment

Due to the lack of depth information in the images, approaches such as [1], [2] and [18] use scale-space representation to localize keypoints at different scales. In their approach, the image is represented by a multilevel, multiscale pyramid in which for each level the image is smoothed and sub-sampled.

Since RGB-D images are composed of color as well as depth information, instead of computing a pyramid and representing the keypoints at the scale-space, we use the depth information of each keypoint to define the scale factor s of the patch to be used in the neighborhood analysis. In this way, patches associated with keypoints farther from the camera will present smaller sizes.

In order to compute the dominant orientation θ for the keypoints, we employ the fast orientation estimator presented in [2]. The orientation assignment for each keypoint is achieved by computing the Haar wavelet responses in both x and y directions. Differently from [2], that uses the scale factor s at which the keypoint was detected to compute the radius ($6s$) of the circular neighborhood around the keypoint, we use the keypoint depth acquired from the RGB-D data. This value is used to scale the size of wavelets ($4s$) and to

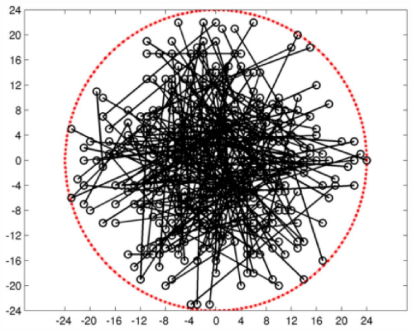


Fig. 3. Patch $\mathbf{p}$ with 48×48 pixels indicating 256 sampled pairs of pixel locations used to construct the local binary pattern.

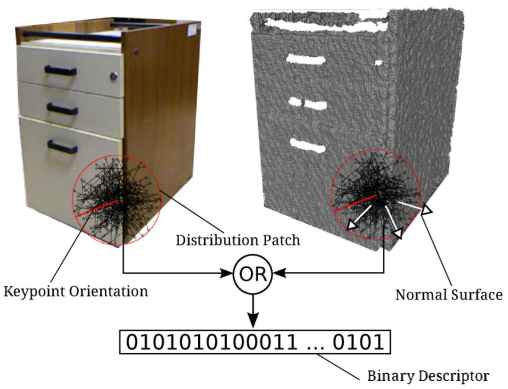


Fig. 4. Diagram of BRAND creation. The patch of size $S \times S$ is centered at the keypoint location. For sampled pair $(\mathbf{x}, \mathbf{y})$, in a patch $\mathbf{p}$, we evaluate changes in the intensity and geometry.

determine the standard deviation $(2s)$ of a Gaussian function used to weight the wavelet.

B. Appearance and Geometry Fusion

There are several choices to compose a descriptor, and bit strings are among the best approaches, mainly due to the reduction in dimensionality and efficiency in computation achieved with their use. In addition, the similarity between two binary descriptors can be measured by the Hamming distance, which can be computed efficiently with a bitwise XOR operation and a bit count.

Although our descriptor encodes point information as a binary string, like approaches described in [5], [18], [6], [7], we embed geometric cues into our descriptor to improve robustness to changes in illumination and the lack of texture in scenes.

BRAND can be formally described as follows. Let the pair (I, D) denote the output of an RGB-D system where, $I(\mathbf{x})$ and $D(\mathbf{x})$ provide, respectively, color and depth information for a pixel $\mathbf{x}$. For spatial points defined by the depth map D, we provide an estimation of their normal vectors as a map N, where $N(\mathbf{x})$ is estimated efficiently by principal component analysis over a small neighborhood in the surface defined by the depth map.

Our descriptor is constructed for a small image patch $\mathbf{p}$, centered at a pixel $\mathbf{k}$. We then use $p_i(\mathbf{x})$ and $p_n(\mathbf{x})$ to denote, respectively, the pixel intensity and surface normal for a pixel $\mathbf{x} \in \mathbf{p}$. The first step to compute the set of descriptors of an RGB-D image (I, D) is the selection of a subset $\mathcal{K}$ of keypoints among image pixels. An efficient keypoint detector, such as [16] or [17] can be used to construct the set $\mathcal{K}$. Indeed, we performed experiments with four different keypoint detectors: [1], [2], [16], [17], and our descriptor presents an average accuracy of 0.57 with standard deviation of 0.035.

Given an image keypoint $\mathbf{k} \in \mathcal{K}$, assume an image patch $\mathbf{p}$ of size $S \times S$ (in this work we consider $9 \leq S \leq 48$) centered at $\mathbf{k}$. Figure 3 illustrates the patch where the set of pixel pairs $(\mathbf{x}_i, \mathbf{y}_i) \in \mathbf{p}$ are indicated with line segments. We use a pattern with locations given by an isotropic Gaussian distribution $\mathcal{N}(0, \frac{48^2}{25})$ for selecting pixel pairs, inspired in the work of [5]. The variance of $\sigma = \frac{1}{25}48^2$ gives best results in terms of recognition rate, according to the experiments performed by Calonder et al. [5]. However, differently from that work, we remove all pairs with points outside of the circle with radius equals to 24. Hence, we guarantee that all pixels within the circle are preserved independent of patch rotation. We also pre-smooth the patch with a Gaussian kernel with $\sigma = 2$ and a window with 9×9 pixels to decrease the sensitivity to noise and increase the stability in the pixels comparison.

Let the set of sampled pairs from $\mathbf{p}$ be denoted by $S = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, 256\}$. In order to guarantee test consistency while searching for correspondences, the same set S is used to construct descriptors for all keypoints sampled from all images.

The patch $\mathbf{p}$ is translated to the origin and then rotated and scaled by the transformation $\mathbf{T}_{\theta,s}$, where θ is the dominant direction and the scale factor s is computed by:

$$s = \max\left(0.2, \frac{3.8 - 0.4\max(2, d)}{3}\right), \tag{1}$$

which linearly scales the radius of circular patch $\mathbf{p}$ from 9 to 48 and filter depths with values less than 2 meters.

To construct our 256 bits feature descriptor we use, from the rotated and scaled patch $\mathbf{p}$, the set:

$$P = \{(\mathbf{T}_{\theta,s}(\mathbf{x}_i), \mathbf{T}_{\theta,s}(\mathbf{y}_i)) | (\mathbf{x}_i, \mathbf{y}_i) \in S\}. \tag{2}$$

Then, we evaluate for each pair $(\mathbf{x}_i, \mathbf{y}_i) \in P$ the function:

$$f(\mathbf{x}_i, \mathbf{y}_i) = \begin{cases} 1 & \text{if } (p_i(\mathbf{x}_i) < p_i(\mathbf{y}_i)) \vee \tau(\mathbf{x}_i, \mathbf{y}_i) \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where the first term captures the characteristic gradient changes in the keypoint neighborhood and $\tau(.)$ function evaluates the geometric pattern on its surface. Figure 4 illustrates the construction process of the bit string.

The analysis of the geometric pattern using $\tau(.)$ is based on two invariant geometric measurements: i) the normal displacement and ii) the surface's convexity. While the normal displacement test is performed to check if the dot product between the normals $p_n(\mathbf{x_i})$ and $p_n(\mathbf{y_i})$ is smaller than a

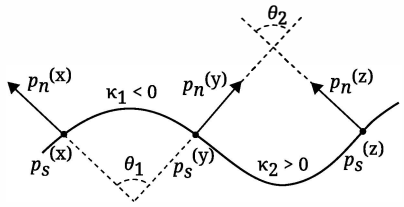
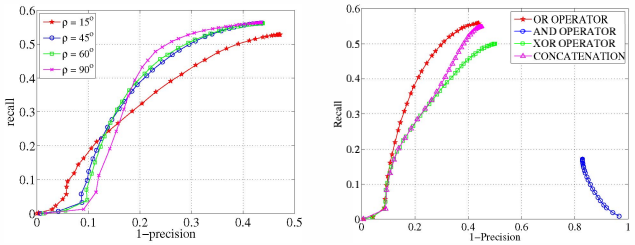


Fig. 5. Example of ambiguity in the dot product. Despite the fact that the points $p_s(\mathbf{x})$ and $p_s(\mathbf{y})$ define a concave surface patch while $p_s(\mathbf{y})$ and $p_s(\mathbf{z})$ define a convex surface patch, the dot products $\langle p_n(\mathbf{x}),p_n(\mathbf{y})\rangle = \langle p_n(\mathbf{y}),p_n(\mathbf{z})\rangle$. In such cases, the curvature signals $\kappa 1 < 0$ and $\kappa 2 > 0$ are used to unambiguously characterize the patch shape.



(a) (b)

Fig. 6. (a) Angular threshold for dot product test. On the average, the best choice is use 45 degrees. (b) The best binary operator to be used for fuse appearance and geometric was the OR operator.

displacement threshold ρ, the convexity test is accomplished by the local curvature signal, κ, estimated as:

$$\kappa(\mathbf{x}_i, \mathbf{y}_i) = \langle p_s(\mathbf{x}_i) - p_s(\mathbf{y}_i), p_n(\mathbf{x}_i) - p_n(\mathbf{y}_i)\rangle, \quad (4)$$

where $\langle . \rangle$ is the dot product and $p_s(\mathbf{x})$ is the $3D$ spatial point associated to the pixel $\mathbf{x}$ and depth $D(\mathbf{x})$. Figure 5 illustrates an example where the dot product between surface normals is ambiguous, since $\theta_1 = \theta_2$, but different signed curvatures, $\kappa_1 < 0$ and $\kappa_2 > 0$, are used to unambiguously characterize these different shapes, besides capturing convexity as additional geometric features.

The final geometric test is given by:

$$\tau(\mathbf{x}_i, \mathbf{y}_i) = (\langle p_n(\mathbf{x}_i), p_n(\mathbf{y}_i)\rangle < \rho) \wedge (\kappa(\mathbf{x}_i, \mathbf{y}_i) < 0)\,. \quad (5)$$

Finally, the descriptor extracted from a patch $\mathbf{p}$ associated with a keypoint $\mathbf{k}$ is encoded as a binary string computed by:

$$b(\mathbf{k}) = \sum_{i=1}^{256} 2^{i-1} f(\mathbf{x}_i, \mathbf{y}_i). \quad (6)$$

In order to show that our descriptor is based on invariant measures, we recall that it combines appearance and geometry. Appearance is an object property invariant to any geometric transformation, although its projection on an image may vary with illumination and other conditions. The geometric component of our descriptor is based on the relation between the normal displacement and the surface convexity, which are geometric measurements invariant to rotation, translation and scaling.

IV. EXPERIMENTS

For evaluation purposes, we perform a set of tests to analyze the behavior of the BRAND descriptor for the matching tasks. Comparisons are performed with the standard approaches of two-dimensional images descriptor, SIFT [1] and SURF [2], with the geometric descriptor, spin-images [3], and the state-of-the-art in fusing both texture and shape information CSHOT [13].

For the experiments, we use the dataset presented in [19]. This dataset is publicly available[1] and contains several real world sequences of RGB-D data captured with a Kinect™sensor. Each sequence in the dataset provides the ground truth of the camera pose estimated by a MoCap system. We selected four sequences in the dataset to use in our experiments: i) *freiburg2_xyz*, in which the Kinect is moving individually along the x/y/z axes; ii) *freiburg2_rpy* with Kinect rotated individually around the three axes; iii) the handheld slam sequence *freiburg2_desk*; and iv) *freiburg2_pionner_slam2* with a Kinect mounted on top of a Pioneer robot.

To evaluate the performance of our descriptor and compare to other approaches, we use the criterion presented in [20]. We match all pairs of keypoints from two different images. If the Euclidean (for SURF and SIFT), Correlation (for spin-image) or Hamming (for BRAND) distance between the descriptors falls below a threshold t, a pair is termed as a valid match. Therefore, we plot the *recall* versus *1-precision* values, obtained by changing the values of t. Recall is the number of correctly matched keypoints and 1-Precision is the number of false matches relative to the total number of matches.

For each sequence, given an RGB-D image of frame i, we compute a set of keypoints $\mathcal{K}_i$ using the STAR detector[2]. To make a fair comparison among all descriptors' approach, the depth information was not used to detect keypoints. All keypoints $\mathbf{k} \in \mathcal{K}_i$ are transformed to frame $i + \Delta$ creating the second set $\mathcal{K}_{i+\Delta}$, using as the ground truth pose those frames ($\mathbf{x_i}$ and $\mathbf{x_{i+\Delta}}$). We compute a descriptor for each keypoint in both sets and then perform the match.

In the following sections, we evaluate several aspects of the proposed descriptor and show comparisons with other methods regarding computation time, memory consumption and accuracy.

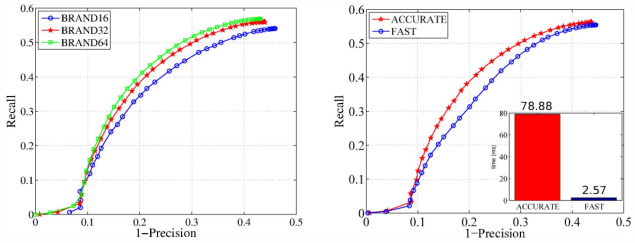
A. Parameter Settings

Experimentally, we found that a threshold ρ that corresponds to 45 degrees for the maximum angular displacement of normals results in a larger number of inliers (Figure 6 (a)).

The plot shown in Figure 7(a) depicts the accuracy *versus* the number of bytes used for the BRAND descriptor. The results show that the accuracy for 32 bytes is similar to the accuracy for 64 bytes. Therefore, to obtain a more compact representation, we have chosen to use 32 bytes in the experiments.

[1]https://cvpr.in.tum.de/data/datasets/rgbd-dataset

[2]STAR detector is an implementation of Center Surrounded Extrema [16] in OpenCV 2.3.



(a) (b)

Fig. 7. (a) Different sizes for the BRAND descriptor; (b) Accurate *versus* Fast normal estimation. Even with the less precise normal estimation, BRAND have high accuracy in the keypoints correspondences.

We chose a bit operator to combine appearance with geometry to maintain the simplicity and computational efficiency of the descriptor. To fuse these information, we evaluated different operators, such as XOR, AND and OR. Figure 6 (a) shows that fusing both texture and geometrical information with OR operator provides a signature with highest discriminative power. Even when compared with concatenation operator. The use of information from two differents domains has disadvantage of being exposed to two different sources of noise. However, using a binary operator rather than concatenation our descriptors are able to balance noise in one domain using other kinds of information.

One problem that arises when using the binary operator OR to set the bits of the descriptors is ambiguity. In general, it may not be known if a bit was set to 1 due to variation in the normal or in intensity. Let D_1 and D_2 be two descriptors with only one bit, and an uniform distribution of the pairs. The following four cases can be listed:

- $D_1 = 0$, $D_2 = 0$: Descriptors are equal, since there is neither normal nor intensity variation reported in both;
- $D_1 = 0$, $D_2 = 1$: Descriptors are different, since there is no variation reported by D_1, but D_2 reports some variation (either normal or intensity);
- $D_1 = 1$, $D_2 = 0$: Descriptors are different, since there is no variation reported D_2, but D_1 reports some variation (either normal or intensity);
- $D_1 = 1$, $D_2 = 1$: Descriptors are equal, since they both report some variation.

In all four cases above, the variation source cannot be known. In the first three cases, the source of variation does not matter, because only one descriptor reports some variation. However, in the last case, both descriptors reports some variation and, if the variation sources were different, the descriptors should not be equal. Hence $D_1 = D_2 = 1$ is an ambiguous case that may happen.

Table I shows all nine cases that can produce $D_1 = 1$ and $D_2 = 1$. In only two of these nine cases the bit was incorrectly set (descriptors D_1 and D_2 should be considered as different, but for this analysis they will be considered as being equal). This occurs when there are changes in the direction of normal but there are no changes in intensity on the surface that generated descriptor D_1, and the surface that produced D_2 does not have variation in the direction of normals, but has changes in intensities. Thus, the probability of comparing

TABLE I

THIS TABLE SHOWS ALL NINE CASES THAT CAN PRODUCE $D_1 = 1$ AND $D_2 = 1$. FOR ALL THESE CASES ONLY TWO CAN BE AMBIGUOUS (COLUMNS 2 AND 4 WITH BITS IN BOLDFACE). CHANGES IN NORMAL OR INTENSITY ARE REPRESENTED WITH BIT EQUAL TO 1.

		1	2	3	4	5	6	7	8	9
$D_1 = 1$	*Normal*	0	**0**	0	**1**	1	1	1	1	1
	Intensity	1	**1**	1	**0**	0	0	1	1	1
$D_2 = 1$	*Normal*	0	**1**	1	**0**	1	1	0	1	1
	Intensity	1	**0**	1	**1**	0	1	1	0	1

ambiguous bits can be estimated as $\frac{1}{4} \times \frac{2}{9} \approx 5.5\%$. In practice, the ambiguity is yet smaller. We computed for 420 keypoints in 300 pairs of images the number of ambiguities and we found the rate to be close to 0.7%.

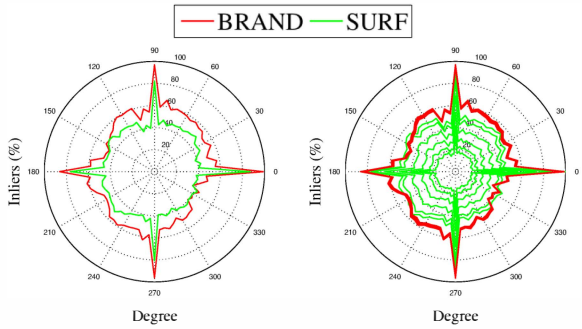
B. Rotation Invariance

We also evaluated the descriptor's invariance to rotation. We use synthetic in-plane rotation and added Gaussian noise with standard deviation equal to 15 degrees (Figure 9 (a)). After applying the rotation and adding noise, we computed the keypoint descriptors using BRAND and SURF, and then performed a brute-force matching to find correspondences.

Figure 9 (b) shows the results for the synthetic test for noise with standard deviation of $15, 30, 45, 60$ and 75. The results are given in terms of percentage of inliers as a function of the rotation degree. Notice that BRAND is more stable and outperforms SURF in all scenarios.

C. Normal Computation

All geometric descriptors used for comparison in the experiments require that point clouds have normals. There are several methods to estimate normals from a point cloud. An accurate approach consists in estimating the surface normal by Principal Component Analysis (PCA) on the nearest neighbors of the keypoint [21]. This was the method used to estimate the normals in all match experiments. However, a less accurate, but faster approach, is to use the pixel neighborhoods defined by the structure from RGB-D images [22]. Figure 7 (b) shows the matching accuracy and the time spent by BRAND



(a) (b)

Fig. 9. Percentage of inliers as a function of rotation degree. (a) BRAND and SURF matching performance under synthetic rotations with Gaussian noise of standard deviation of 15; (b) BRAND and SURF matching sensitivity under 0, 15, 30, 45, 60 and 75 levels of noise. BRAND is virtually unaffected.

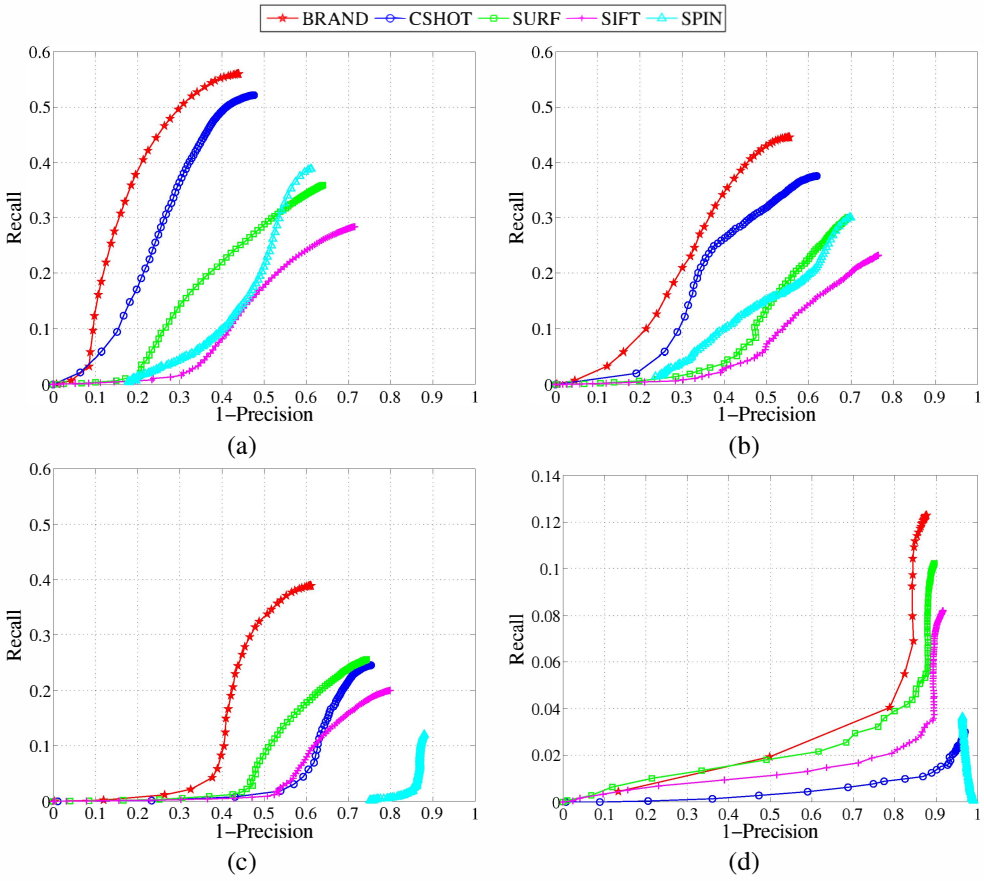


Fig. 8. Precision-Recall curves for (a) freiburg2_xyz, (b) freiburg2_rpy, (c) freiburg2_desk and (d) freiburg2_pionner_slam2. The keypoints were detected using STAR detector. BRAND outperforms all others approaches, including CSHOT, which combines, like BRAND, texture and geometric information.

using both estimating techniques. We can see that, even with a less precise normal estimation, BRAND presents higher accuracy in the correspondences. Therefore, BRAND can be optimized if necessary for a given application without penalizing significantly its accuracy.

D. Comparisons

We have recorded the creation and matching time. The experiments were executed in an Intel Core i5 2.53GHz (using only one core) running Ubuntu 11.04 (64 bit). The values were averaged over 300 runs and all keypoints were detected by the STAR detector. We clearly see in Figure 10(a) that

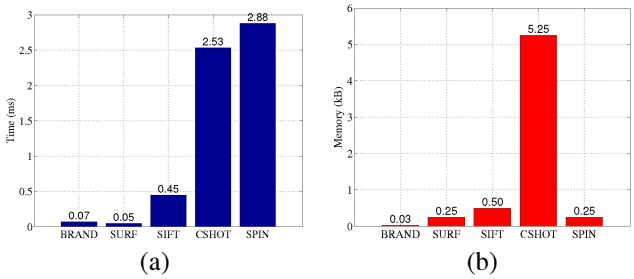


Fig. 10. Comparison between descriptors using: (a) processing time to create a keypoint descriptor and (b) the memory consumption in kbytes of each descriptor.

BRAND is faster than the other descriptors in the creation step, losing only for SURF. Addionally, BRAND presents the lowest memory consumption with 32 bytes for keypoint descriptors, while CSHOT, like BRAND, which combines appearance and geometry, has descriptors of 5.25 kBytes in size (Figure 10(b)).

Figure 8 shows the results of the threshold-based similarity matching tests. As illustrated in the precision-recall curves, the BRAND descriptor showed a significantly better performance than all other approaches in all sequences. Even for the two more challenging sequences, *freiburg2_desk* and *freiburg2_pionner_slam2*, which have high camera speed, and also in the particular case of *freiburg2_pionner_slam2* sequence, with few (which was acquired with the robot joysticked through a textureless large hall).

E. Alignment Quality

We also examined the performance of our descriptor on the registration task. Usually, a registration algorithm is divided in two main steps: coarse and fine alignment. We use BRAND descriptor in the coarse alignment to compute an initial estimation of the rigid motion between two clouds of 3D points using correspondences. In the fine alignment, we use Iterative Closest Point (ICP) algorithm to find a local optimum solution based on the prior coarse alignment with BRAND.

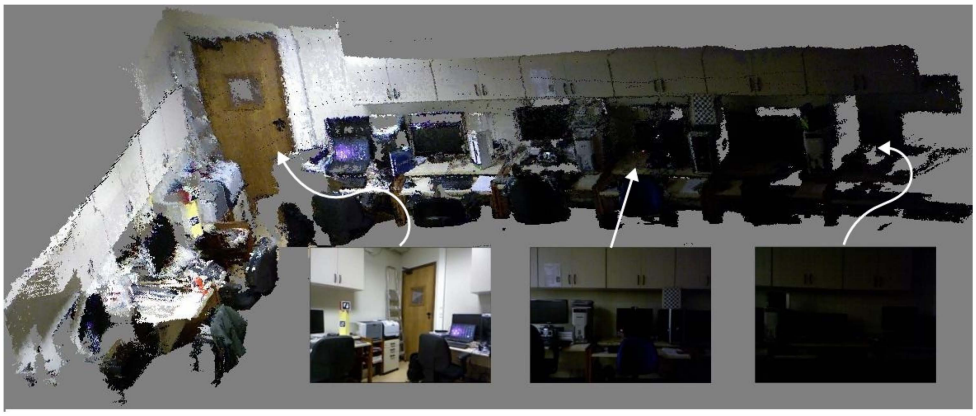


Fig. 11. Registration of a partially illuminated lab. The frames were used with images from a scene ranging from well illuminated to complete darkness. As BRAND contains geometric information, it is possible to match even if the scene is under inadequate illumination.

The tests were performed in a room with poor lighting to show that we can register the clouds even with sparsely illuminated environments since the BRAND descriptor also contain geometric information. Due to the lack of RGB information in the regions without illumination, we implemented an alignment algorithm which uses the geometrical keypoint detector NARF [23] whenever the number of STAR keypoints is below a threshold. We have acquired several frames from our lab with regions ranging from well illuminated to completely dark. The final alignment is shown in Figure 11. This result makes it clear that, even for some regions without illumination, it was possible to successfully accomplished alignment of the point clouds.

V. CONCLUSIONS

We proposed a new descriptor named BRAND. This descriptor takes into account appearance and geometry from RGB-D images, presenting orientation invariance and robustness to different illumination conditions. In our experiments, BRAND outperformed all the other descriptors, including the state of the art CSHOT descriptor, which also fuses appearance and geometry. Experiments demonstrate that our technique is robust for registration tasks under poor lightning and sparsely textured scenes.

The results presented here extend the conclusion of [4], [13], [24] where the arrangement of appearance and geometric information is advantageous not only in perception tasks, but also useful to improve the quality of the correspondence process. Appearance and geometry information indeed enable better performance than using either information alone.

The main constraint of our methodology are the bumpy surfaces. Since the geometrical features are extracted using a threshold for the displacement between normals, the small regularities of these surfaces can be confused with noise.

REFERENCES

[1] D. G. Lowe., "Distinctive image features from scale-invariant keypoints," *IJCV*, pp. 91–110, 2004.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," *Proc. CVIU*, vol. 110, pp. 346–359, June 2008.

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

[4] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *ICRA*, 2011.

[5] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *ECCV*, September 2010.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *ICCV*, Barcelona, November 2011.

[7] M. Ambai and Y. Yoshida, "CARD: Compact And Real-time Descriptors," in *ICCV*, Barcelona, November 2011.

[8] J. Choi, W. R. Schwartz, H. Guo, and L. S. Davis, "A Complementary Local Feature Descriptor for Face Identification," in *WACV*, 2012.

[9] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.

[10] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *ICRA*, 2009, pp. 1848–1853.

[11] R. B. Rusu, Z. C. Marton, N. Blodow, and M. Beetz, "Persistent Point Feature Histograms for 3D Point Clouds," in *Proc. of International Conference on Intelligent Autonomous Systems (IAS-10)*, 2008.

[12] A. Zaharescu, E. Boyer, K. Varanasi, and R. P. Horaud, "Surface Feature Detection and Description with Applications to Mesh Matching," in *CVPR*, Miami Beach, Florida, June 2009.

[13] F. Tombari, S. Salti, and L. D. Stefano, "A combined texture-shape descriptor for enhanced 3D feature matching," in *ICIP*, 2011.

[14] ——, "Unique Signatures of Histograms for Local Surface Description," in *ECCV*, 2010.

[15] A. Kanezaki, Z.-C. Marton, D. Pangercic, T. Harada, Y. Kuniyoshi, and M. Beetz, "Voxelized Shape and Color Histograms for RGB-D," in *IROS Workshop on Active Semantic Perception*, September 2011.

[16] M. Agrawal, K. Konolige, and M. R. Blas, "CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching," in *Proc. ECCV*, 2008.

[17] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *PAMI*, pp. 105–119, 2010.

[18] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *ICCV*, 2011.

[19] J. Sturm, S. Magnenat, N. Engelhard, F. Pomerleau, F. Colas, W. Burgard, D. Cremers, and R. Siegwart, "Towards a benchmark for RGB-D SLAM evaluation," in *Proc. of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at RSS*, June 2011.

[20] Y. Ke and R. Sukthankar, "PCA-SIFT: A More distinctive Representation for Local Image Descriptors," in *CVPR*, 2004.

[21] 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.

[22] D. Holz, S. Holzer, R. B. Rusu, and S. Behnke, "Real-Time Plane Segmentation using RGB-D Cameras," in *RoboCup Symposium*, 2011.

[23] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Point Feature Extraction on 3D Range Scans Taking into Account object boundaries," in *Proc. ICRA*, May 2011.

[24] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments," in *ISER*, 2010.