# Appearance and Geometry Fusion for Enhanced Dense 3D Alignment

Erickson R. Nascimento William Robson Schwartz Gabriel L. Oliveira
Antônio W. Veira Mario F. M. Campos Daniel B. Mesquita
*Departamento de Ciência da Computação*
*Universidade Federal de Minas Gerais*
*Belo Horizonte, Brazil*
*Email: {erickson, william, gabriel, awilson, mario, balbino}@dcc.ufmg.br*

Fig. 1. Registration of a partially illuminated scene, a hard task due to the lack of textural information. The cloud alignment was done using the proposed RGB-D descriptor, *Binary Appearance and Shape Elements* (BASE), that combines appearance and shape.

***Abstract*—This work proposes a novel RGB-D feature descriptor called Binary Appearance and Shape Elements (BASE) that efficiently combines intensity and shape information to improve the discriminative power and enable an enhanced and faster matching process. The new descriptor is used to align a set of RGB point clouds to generate dense three dimensional models of indoor environments. We compare the performance of state-of-the-art feature descriptors with the proposed descriptor for scene alignment through the registration of multiple indoor textured depth maps. Experimental results show that the proposed descriptor outperforms the other approaches in computational cost, memory consumption and match quality. Additionally, experiments based on cloud alignment show that the BASE descriptor is suitable to be used in the registration of RBG-D data even when the environment is partially illuminated.**

***Keywords*-Descriptor, RGB-D Camera, Reconstruction, Registration, Three-dimensional Mapping**

## I. Introduction

Building accurate 3D models of a scene is a fundamental problem in Computer Graphics and Computer Vision. Methodologies to build these models usually face the task of alignment and registration, which are related to finding an affine transformation T between two different views of the scene in order to represent both in a single coordinate system. The approaches to carry the task of alignment usually employ some kind of descriptors in order to establish a set of corresponding points between two views that will be used to find an approximation for T. Hence, constructing descriptors able to correctly establish pairs of such corresponding points is of central importance for alignment and, consequently, for 3D registration as well.

The approaches for extracting descriptors can be categorized according to the nature of data acquired to represent a view of a scene. For instance, data may be textured images or depth images.

Textured images are popular and provide such a rich source of information that naturally pushed the use of texture-based descriptors in several methods for alignment despite the inherent complexity involved. Therefore, alignment and registration from textured images has become one of the fundamental issues in Computer Vision and Robotics, and are at the heart of important tasks such as tracking and Simultaneous Localization And Mapping (SLAM). Computer Vision literature presents numerous works on using different cues for correspondence based on texture, such as Scale Invariant Feature Descriptor (SIFT) [1], Speeded-Up Robust Descriptor (SURF) [2], Binary Robust Independent Elementary Features (BRIEF) [3], Binary Robust Invariant Scalable Keypoints (BRISK) [4] and Oriented FAST and Rotated BRIEF (ORB) [5]. In virtually all these approaches, feature

descriptors are estimated from images alone, and they rarely use other information such as geometry. As a consequence, common issues concerning real scenes such as variation in scene illumination and textureless objects may dramatically decrease the performance of the texture-based descriptors.

Depth images, although with their increasing use, are less popular and the geometrical nature of the data involves higher complexity to define descriptors and usually have large ambiguous regions which does not allow correspondence. To define robust descriptors for geometrical data, large amount of data is necessary to encompass enough information to avoid ambiguities. Spin-Image [6], Fast Point Feature Histograms (FPFH) [7], Normal Aligned Radial Feature (NARF) [8], Point Feature Histograms (PFH) [9] are some examples of such descriptors. Even though these descriptors can handle textureless scene regions, where texture based descriptors fail, their construction involves complex geometrical operations, resulting high processing time and memory consumption.

The combination of appearance and three-dimensional shape cues is still in its prelude. However, as far as accuracy is concerned, Lai et. al. [10] have already shown that the combined use of intensity and depth outperforms view-based distance learning using either intensity or depth alone. Additionally, Zaharescu et. al [11] and Tombari et.al [12] have shown that the use of features of differents domains is a very promising approach to improve the quality in matching task in the descriptor level.

With the recent introduction of fast and inexpensive RGB-D sensors (where *RGB* implies trichromatic intensity information and *D* stands for depth), the integration of synchronized intensity (color) and depth has become feasible. RGB-D systems output color images and the corresponding pixel depth information, which enable the acquisition of both depth and visual cues in real-time. These systems have opened the opportunity to obtain 3D information with unprecedented richness. One such system is the Kinect$^{TM}$sensor [13], a low cost commercially available system that produces RGB-D data in real-time for gaming applications. Given the technological advances of RGB-D sensors and the use of large data sets, fast and low memory consumption descriptors that efficiently use the available information, play a central role in a myriad of tasks, such as, 3D modeling, registration, surface reconstruction, object detection and recognition systems for mapping tasks.

In this paper, we propose a novel RGB-D feature descriptor called Binary Appearance and Shape Elements (BASE) that efficiently combines intensity and shape information to improve the discriminative power providing enhanced and faster matching process. Experimental results demonstrate that the proposed descriptor outperforms the state-of-the-art feature descriptors and provides indoors 3D scene alignment with the smallest error.

After discussing related works in the next section, in Section III we present the proposed descriptor and then the RGB-D point cloud registration process used in the work. Experiments are presented in Section IV followed, in Section V, by the conclusions we have reached with this investigation.

## II. Related Work

A great challenge for registering multiple depth maps is related to the process of recovering the rigid affine transformation $T$ to describe two depth maps into a single coordinate system. To address this issue, descriptors have been applied to find corresponding points from two depth maps in order to constrain the search space for the transformation $T$. The work proposed by Vieira et al. [14] uses a descriptor to propose an iterative framework to address pair-wise alignment of a sequence of depth maps while ensuring global coherence of the registration for implicit reconstruction purpose. A global alignment algorithm that does not use local feature descriptors was presented by [15] using Extended Gaussian Images.

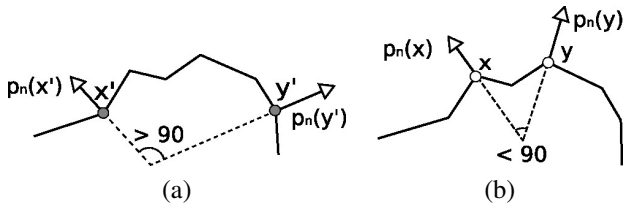
Independently of strategies used to pre-align depth maps, a common requirement is that data have sufficient overlap in order to establish correspondences and a graph defining which pairs, among all depth maps, have such overlap. Most commercial packages, such as [16], requires that users select manually the pairs to be aligned. Furthermore, this pre-alignment is generally refined by local minimization algorithms, such as the classical Iterative Closest Point (ICP) [17] in order to achieve the best alignment, given an initial guess of pre-alignment.

Non-rigid and scale invariant registration such as proposed in [18] and [19] are most used for matching purpose rather than reconstruction. A survey on range image registration has been presented in [20], where different methods for pre-alignment and fine registration are compared in terms of robustness and efficiency.

In the field of image processing, SIFT [1], SURF [2] are the most used algorithms for keypoint extraction and descriptor creation. These methodologies build their feature detectors and descriptors based on local gradients and specific orientation to achieve rotational invariance. Inspired by the idea of Local Binary Patterns (LBP) [21], works such as [3], [4], [5], [22] presented a new family of descriptors that use binary strings to build a descriptor. This approach for building descriptors presents the advantage of small memory usage and low processing time.

Feature extraction from 3D data has been successfully obtained with the *spin-image* [6], which creates a 2D representation of the surface patch surrounding a 3D point. Object edges constitute an important challenge that has been tackled by another descriptor for 3D point clouds known as NARF [8], which identifies edges of objects based on transitions between foreground and background. Others approaches proposed to handle point clouds are [7], [9].

If on one hand texture information on an image can usually provide better perception of object features, on the other hand depth information produced by 3D sensors is less sensitive to lighting conditions. Recently, several descriptors have been proposed to combine multiple cues. Kanezaki et al. [23] presented the Voxelized Shape and Color Histograms (VOSCH) descriptor, which by combining depth and texture, can increase the recognition rate in cluttered scenes with




(a) (b)

Fig. 2. Image (a) shows a surface where points $\mathbf{x}'$ and $\mathbf{y}'$ have normals with displacement greater than 90 degree leading to bit value 1. In image (b) is shown the normals of the points $\mathbf{x}$ and $\mathbf{y}$ that lead to bit 0 due to a displacement less than 90 degree.

obstruction. However, different from our approach, VOSCH is a global descriptor. In [11] the authors present the Mesh-HOG descriptor. This descriptor uses texture information of 3D models as scalar functions defined over a 2D manifold. Tombari et al. [12] proposed the descriptor called Color-SHOT (CSHOT) based on an extension of their shape only descriptor Signature of Histograms of Orientations (SHOT) [24] to incorporate texture. CSHOT descriptor combines two histograms, one with the geometric features over the spherical support around the keypoint and the other containing the sum of the absolute differences between the RGB triples of the each of its neighboring points. CSHOT is compared against MeshHOG in [12] and the authors reported that CSHOT outperformed MeshHOG in processing time and accuracy.

Similar to CSHOT and MeshHOG, our descriptor is a local descriptor and brings forth the advantages of both texture and depth. However, unlike these descriptors our approach uses smaller memory space and is faster without losing the discriminative power, as it will be shown in the experimental results.

## III. Methodology

In this section we detail the design of our novel feature descriptor and also describe the method employed to perform the registration of multiple indoor textured depth maps.

Unlike traditional approaches used in the last years that employ only texture information as [1], [2], [25], [3] or shape [6], [8], the keypoint descriptor developed in this work encodes geometrical and appearance information simultaneously.

### A. BASE Descriptor

In order to detail our descriptor, let $M = \{1, 2, \ldots, m\}$, $N = \{1, 2, \ldots, n\}$ and let us denote the output of a RGB-D camera as a pair $(I, D)$ where

$$I : M \times N \to C$$

maps each image pixel $\mathbf{x} = (i, j)$ of our $m \times n$ image to an intensity $c = I(\mathbf{x}) \in C$ where $C = \{0, \ldots, 255\}$ (we consider only the intensity and not the color information), and

$$D : M \times N \to \mathbb{R}^+$$

maps each image pixel $\mathbf{x}$ to its depth value $d = D(\mathbf{x}) \in \mathbb{R}^+$.

For each spatial point defined by the depth map $D$, we provide an estimation of its normal vector as a map

$$V : M \times N \to \mathbb{R}^3$$

where the vector $v = V(\mathbf{x}) \in \mathbb{R}^3$ is estimated using a small neighborhood in the surface defined by the depth map.

The first step to compute the set of descriptors for an RGB-D image $(I, D)$ is the selection of a subset $\mathcal{K} \subset M \times N$ of keypoints $\mathbf{k}$ among the image pixels. We use an efficient keypoint detector called CenSurE [26] to construct our set $\mathcal{K}$.

Given an image keypoint $\mathbf{k} \in \mathcal{K}$, we consider an image patch $\mathbf{p}$ with $S \times S$ pixels centered at $\mathbf{k}$ and define the map

$$p_i : \{1, \ldots, S\} \times \{1, \ldots, S\} \to C$$

where $p_i(\mathbf{x}) = I(\mathbf{k} + \mathbf{x} - \mathbf{s})$ and $\mathbf{s}$ is the central pixel of patch $\mathbf{p}$, to map a pixel from local coordinate system of $\mathbf{p}$ to global coordinate system of the image and

$$p_n : \{1, \ldots, S\} \times \{1, \ldots, S\} \to \mathbb{R}^3$$

where $p_n(\mathbf{x}) = N(\mathbf{k} + \mathbf{x} - \mathbf{s})$ to map a pixel from $\mathbf{p}$ to the normal vector of its corresponding position on the image.

To construct our 256 bits feature descriptor we sample a set $P = \{(x_i, y_i), i = 1, \ldots, 256\}$ with 256 pairs of pixel locations from the patch $\mathbf{p}$. This set $P$ is fixed and used to construct descriptors for all keypoints sampled from all images. Fig. 3 illustrates a patch where the set of pixel pairs is indicated with line segments. We then evaluate, for each pair $(x, y) \in P$, the function:

$$f(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 & \text{if } p_i(\mathbf{x}) < p_i(\mathbf{y}) \vee \langle p_n(\mathbf{x}), p_n(\mathbf{y}) \rangle \leq \rho \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\langle p_n(\mathbf{x}), p_n(\mathbf{y}) \rangle$ is the dot product between the point normals $p_n(\mathbf{x})$ and $p_n(\mathbf{y})$, which captures the normal displacements, ranging from $\rho = -1$ to $\rho = 1$.

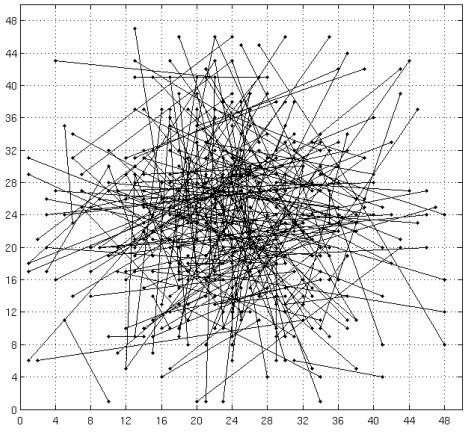



Fig. 3. Patch $\mathbf{p}$ with $48 \times 48$ pixels indicating 256 sampled pairs of pixel locations used to construct the binary feature.

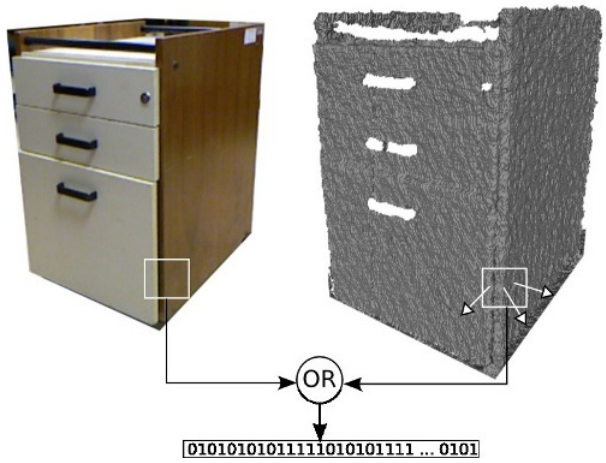



Fig. 4. Creation Diagram of BASE Descriptor. A patch of size $S \times S$ centered at the location of each keypoint in $\mathcal{K}$. For all positions in a set of $(\mathbf{x}, \mathbf{y})$-locations is evaluated the intensity changes in image and degree among the normal points inside of projected patch in the point cloud.

The function $f(\mathbf{x}, \mathbf{y})$ extracts the visual and geometrical features and combines them in a unique vector which represents the signature of a keypoint. The visual feature extraction is based on the direction of the gradient around a keypoint. The idea behind this step is similar to the one used by the Local Binary Patterns (LBP) [21]. The geometrical features depend on normals surface displacement. Figure 2 illustrates two possible cases of normal displacement from a pair $(\mathbf{x}, \mathbf{y})$.

The final descriptor to the patch $\mathbf{p}$ is encoded as a binary string computed by:

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

Figure 4 illustrates the whole process for constructing our descriptor to encode geometrical and appearance information.

As suggested by Calonder et al. [3], we use an image patch of size $S = 48$. After several experiments, we defined the threshold $\rho = 0$ that lead to 90 degrees for the maximum displacement of normals. As in [3], we pre-smooth the patch with a Gaussian kernel with $\sigma = 2$ and window with $9 \times 9$ pixels and, finally, the set of tests locations $(\mathbf{x}_i, \mathbf{y}_i)$ were sampled from an isotropic Gaussian distribution $\mathcal{N}(0, \frac{S^2}{25})$.

### *B. RGB-D Point Cloud Registration Approach*

The main goal of the registration process is to find an affine transformation $T$ between two point clouds taken from different view positions.

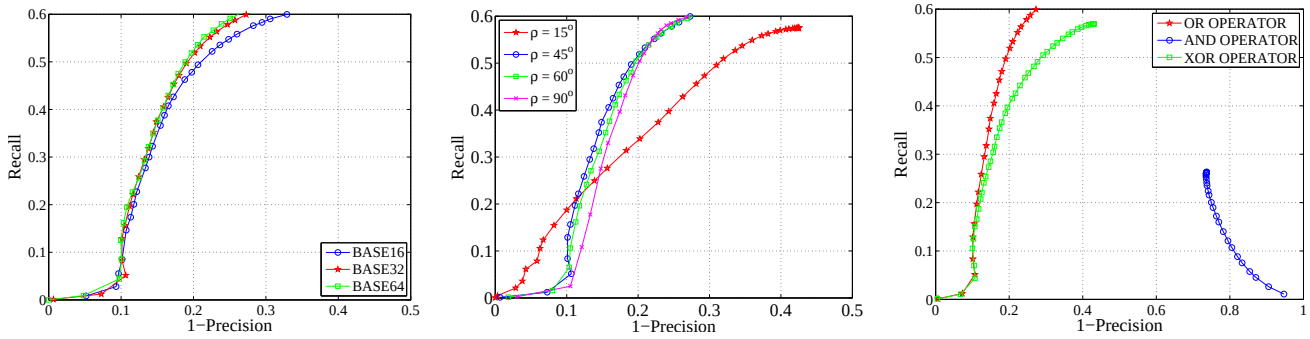
The approach used to register point clouds in this work is divided in two steps: Coarse and fine alignment. In the coarse alignment, we compute an initial estimation $T$ of the rigid motion between two clouds of 3D points using correspondences provided by a feature descriptor. Then, in the fine alignment, we employ the ICP algorithm to find a local optimum solution based on the prior coarse alignment. The ICP algorithm uses an initial estimate of the alignment and then refine the transformation matrix $T^*$ by minimizing the distances between the closest points. The ICP was considered due to its simplicity and low computational time.

The registration process is summarized in the Algorithm 1. It has four main steps:

1) **Keypoint Descriptors**: The function $\mathrm{ExtractDescriptor}$ receives point clouds source and target, denoted by $\mathcal{P}_s$ and $\mathcal{P}_t$, respectively, and returns corresponding sets of keypoints with their descriptors, denoted by $\mathcal{K}_s$ and $\mathcal{K}_t$. The first step to compute the set of descriptors for an image or, in our case, a RGB-D point cloud, is to select a subset of points, called keypoints. A judicious selection of points with property like repeatability provides good detection from multiple views and allows constrained search space for features making the registration suitable to online applications.
2) **Matching Features**: The function $\mathrm{matchDescriptor}$ matches two set of descriptors, $\mathcal{K}_s$ and $\mathcal{K}_t$, to return a set $\mathcal{M}$ of correspondence pairs among source and target point clouds. The distance metric used varies with the type of feature descriptor used. The BASE descriptor considers the *Hamming* distance metric. One of the greatest advantages of using binary string as descriptors, besides its simplicity, is its low computational cost and memory consumption, whereas each descriptor comparison can be performed using a small number of instruction on modern processors. For instance, modern architectures have only one instruction (POPCNT) to count the number of bit sets in a bit vector [27].
3) **Coarse Alignment with SAC**: The function $\mathrm{coarseAlignmentSAC}$ is used to provide an initial transformation $\mathrm{T}$ using the matching set $\mathcal{M}$. We used a Sampled Consensus-Initial Alignment (SAC) approach [28] to reduce the outliers in correspondences (false correspondences). The initial transformation $\mathrm{T}$ is usually not accurate but constrains to a local search for the optimal transformation using a fine alignment algorithm. We noted, as expected, that less descriptive features provide smaller set of inliers than

**Algorithm 1** Point Cloud Alignment($\mathcal{P}_s$, $\mathcal{P}_t$)

1: $(\mathcal{K}_s, \mathcal{K}_t) \leftarrow \mathrm{ExtractDescriptor}(\mathcal{P}_\mathrm{s}, \mathcal{P}_\mathrm{t})$
2: $\mathcal{M} \leftarrow \mathrm{matchDescriptor}(\mathcal{K}_s, \mathcal{K}_t)$
3: $\mathrm{R} \leftarrow \mathrm{coarseAlignmentSAC}(\mathcal{M})$
4: **repeat**
5: $\quad \mathcal{A} \leftarrow \mathrm{closestPoints}(\mathcal{P}_s, \mathrm{R}(\mathcal{P}_t))$
6: $\quad$ Find $\mathrm{T}$ solving:
7: $$\mathrm{T} \leftarrow \arg\min_{T^*} \frac{1}{|\mathcal{A}|} \sum_{(p_s, p_t) \in \mathcal{A}} |p_s - T^*(p_t)|^2$$
8: $\quad \mathrm{R} \leftarrow \mathrm{T} \times \mathrm{R}$
9: **until** MaxIter Reached **or** ErrorChange(T) $\leq \theta$




(a) (b) (c)

Fig. 5. (a) Different sizes for the BASE descriptor; (b) Angular threshold for dot product test. On the average, the best choice is to use 45 degrees; (c) The best binary operator to be used to fuse appearance and geometry was $OR$ operator.

more descriptive features.

4) **Fine Alignment**: Finally, the function closestPoints receives the pre-aligned sets $\mathcal{P}_s$ and $\mathcal{P}_t$ and, constructs the set $\mathcal{A}$ of pairs. The set of pre-aligned pairs $\mathcal{A}$ is then used to find a refined transformation in an iterative process. We use a kd-tree for finding the closest point and, differently from the work by Henry et al. [29] which minimizes a non-linear error, we choose an ICP variant that minimizes the error function point-to-point $\sum |p_s - \mathrm{T}(p_t)|^2$. This error function can be solved using the Horn closed-form [30].

## IV. Experiments

To evaluate the performance of the proposed descriptor, we initially perform a set of tests to evaluate the behavior of our descriptor for matching tasks. Then, we examine its performance, accuracy and robustness for the registration task.

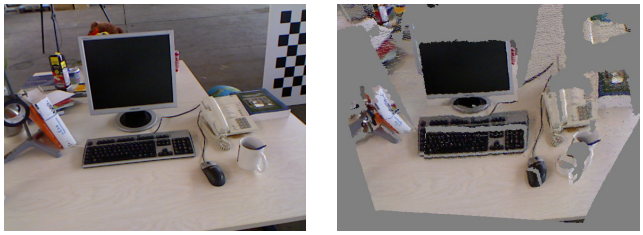
In the experiments, we use the public dataset presented in [31], which is available for download in[1]. This dataset contains several real world sequences of RGB-D data captured with a Kinect™ sensor. The images were acquired at frame rate of 30 Hz and resolution of $640 \times 480$ pixels. Figure 7 shows a frame of two sequences in Freiburg dataset. Each sequence in the dataset provides the ground truth of the camera pose estimated by a MoCap system.

Among the sequences in the dataset, we select two of them to use in our experiments:

- *freiburg2_xyz*: In this sequence the Kinect is moving individually along the x/y/z axes;
- *freiburg2_rpy*: The Kinect was rotated individually around the three axes.

In each sequence, given an RGB-D image of the i-$th$ frame, we compute a set of keypoints $\mathcal{K}_i$. 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 these frames ($\mathbf{x_i}$ and $\mathbf{x_{i+\Delta}}$). We compute a descriptor for each keypoint in both sets and then match them.

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



RGB Image Point Cloud

Fig. 7. RGB Image and Point Cloud example of a frame from freiburg_xyz sequence used in matching experiments.

We use the same criterion presented in [32] and [33] to evaluate the matching performance of the descriptors. First, we detect a set of keypoints using STAR detector[2]. Then, we match all pairs of keypoints from two different RGB-D images. If the Euclidean (for SURF and SIFT), Correlation (for spin-image), dot product (for CSHOT) or Hamming (for BASE) distance between the descriptors falls below a threshold $t$, a pair is considered a match. This threshold is changed to create the *recall* versus *1-precision* curves.

To compute the *recall* and *1-precision*, we count the number of correct matches, termed *true positive*, and the number of incorrect matches, called *false positive*. The *recall* values are determined by:

$$\textit{recall} = \frac{\#truepositive}{\#correspondences},$$

where *#correspondences* is the number of existing correspondences in both images. The *1-precision* values express the number of false detections relative to the total number of detection and it is computed using:

$$\textit{1-precision} = \frac{\#falsepositive}{\#truepositive + \#falsepositive}.$$

[2]STAR detector is a implementation of Center Surrounded Extrema [26] in OpenCV 2.3.1.

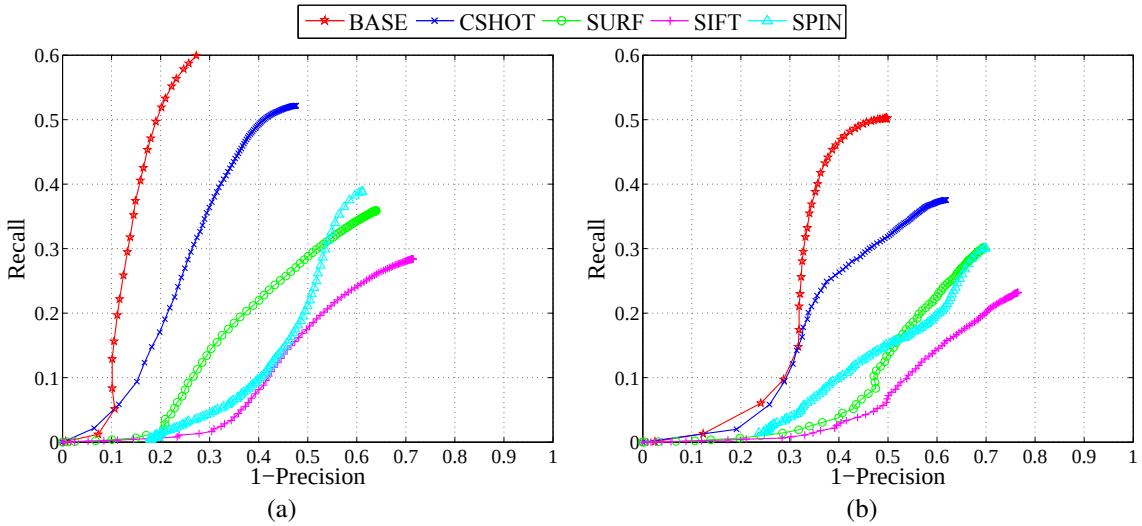



Fig. 6. Precision-Recall curves for (a) freiburg2_xyz, (b) *freiburg2_rpy*. The keypoints were detected using STAR detector [26]. The BASE descriptor outperforms all others approaches, including the state-of-the-art CSHOT, which combined visual and shape information.

### A. Parameter Analysis

We analyze experimentally the best values for the parameters: i) angular threshold; ii) descriptor size and iii) binary operator. We can see in the plots shown in Figure 5 that the best option is the combination of a angular threshold of 45 degrees and the $OR$ operator. Furthermore, we chose the size of 32 bytes as default size since the accuracy for 32 bytes and 64 bytes are similar.

Despite the high quality of matching when using $OR$ binary operator, the fusion using this operator may create ambiguity in the final descriptor. Thus, bits set to 1 due to variation in the normal or intensity may be not distinguishable. However, there exists a small probability of such ambiguity, as described as follows.

Consider strings $L$ and $R$, and their bits $l_i$ and $r_i$ $(i = 1, \ldots, 256)$ to be compared, and an uniform distribution of the pairs. We have four cases from which only one leads to ambiguity:

- $l_i = 0$ and $r_i = 0$: there is no ambiguity because neither the intensity nor the normal varies.
- $l_i = 0$ and $r_i = 1$: there is no ambiguity because there was no variation on the left patch and there was some (intensity or normal variation) on the right patch.
- $l_i = 1$ and $r_i = 0$: there is no ambiguity because there was no variation on the right patch and there was some variation on the left patch.
- $l_i = 1$ and $r_i = 1$: ambiguity may exist. There are nine different situations that can lead to this configuration. Among them, only two can actually generate ambiguity.

Hence, there is only (1/4)*(2/9) = 0.05 (5%) probability of ambiguity per bit.

### B. Matching Performance

To analyze the capability of the BASE descriptor in the matching task, the performance of our descriptor was compared with the standard approaches for two-dimensional image description, SIFT [1] and SURF [2], with the geometric descriptor spin-images [34], and the state-of-the-art descriptor in fusing texture and shape information CSHOT [12].

Figure 6 shows the *recall* vs. *1-precision* curves for each algorithm. We can readly see that, for both sequences, BASE descriptor outperformed all the others approaches, including the state-of-the-art CSHOT.

### C. Time and Memory Consumption

We have recorded the creation time for each descriptor. The experiments were executed in an Intel Core i5 2.53GHz (using only one core). The values were averaged over 300 runs and all keypoints were detected by the STAR detector. We clearly see in Figures 9 and 8 that BASE outperforms the other descriptors in the processing time and memory consumption. Our descriptor presents the lowest memory consumption with 32 bytes for keypoint descriptors, while the state-of-the-art CSHOT, which combines appearance and geometry, has descriptors of 5.25 kBytes in size (Figure 8).

### D. Registration Results

Finally, we examine the performance of our descriptor to the registration task for several images of a research laboratory collected with a Kinect sensor (see Figure 10 and the teaser). We create five challenging sets with different views:

1) Lab180: point cloud with holes (regions not seen by the sensor);
2) Boxes: scene with three object (boxes) with similar geometry;
3) Robots: scene with three robots with the same geometry and texture;
4) Wall: scene rich with textureless regions and
5) Teaser (Figure 1): a set of point clouds acquired from a partially illuminated scene.

TABLE I

THIS TABLE SHOWS MEAN VALUES OF THE ICP ERROR, NUMBER OF INLIERS RETAINED BY SAC IN THE COARSE ALIGNMENT AND TIME SPENT TO REGISTER TWO CLOUDS. IN ALL EXPERIMENTS, THE USE OF OUR DESCRIPTOR (BASE) SPENT LESS TIME AND PROVIDED SMALLER ERROR OF ICP (WHICH INDICATES A BETTER ALIGNMENT) THAN OTHER DESCRIPTORS.

| | Robots (41 frames) | | | Boxes (58 frames) | | | Lab180 (67 frames) | | | Wall (131 frames) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Descriptor** | **Score** | **#Inliers** | **Time (s)** | **Score** | **#Inliers** | **Time (s)** | **Score** | **#Inliers** | **Time (s)** | **Score** | **#Inliers** | **Time (s)** |
| BASE | **0.0025** | 116.95 | **0.30** | **0.0002** | **108.96** | **0.27** | **0.0041** | 53.00 | **0.68** | **0.0001** | 70.96 | **0.71** |
| SURF | 0.0035 | 96.59 | 0.69 | **0.0002** | 58.39 | 0.31 | 0.0070 | 82.09 | 2.40 | 0.0004 | 46.47 | 0.97 |
| SIFT | 0.0058 | 152.10 | 1.28 | 0.0042 | 99.52 | 1.24 | 0.0281 | 129.23 | 6.29 | 0.0021 | 69.66 | 2.09 |
| SPIN | 0.0046 | **155.05** | 2.56 | 0.0017 | 71.30 | 1.70 | 0.0356 | **176.82** | 8.13 | 0.0205 | **181.60** | 9.18 |
| CSHOT | 0.0043 | 143.49 | 2.29 | 0.0002 | 53.54 | 1.30 | 0.0095 | 113.52 | 2.60 | 0.0013 | 66.29 | 2.40 |

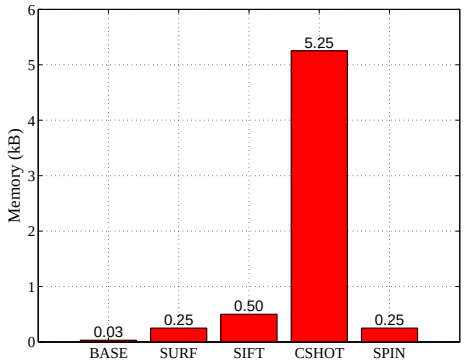



Fig. 8. Comparison between descriptors usin the memory consumption in kbytes of each descriptor. BASE uses only 32 bytes of memory, while SURF and Spin-Image use 256 bytes, SIFT uses 512 bytes and CSHOT 5376 bytes.

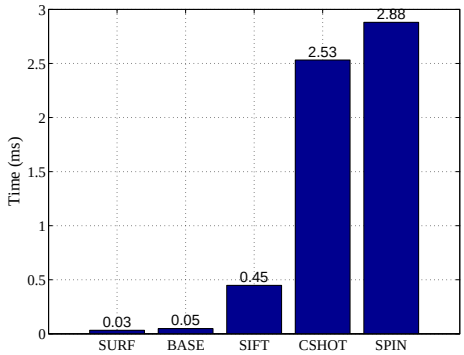



Fig. 9. Comparison between descriptors regarding the processing time to create a keypoint descriptor.

The experiments were performed on a computer running Linux on an Intel core i5 with 4 Gb of RAM. For each final alignment we evaluated the alignment error returned by ICP, the number of inliers retained in the coarse alignment and the time spent for fine and coarse alignment.

Table I shows the registration results. We note that the alignment with the BASE descriptor provides the smaller error despite of its low computational. Figure 10 shows visual results of the alignment achieved using BASE.

As the BASE descriptor considers shape information and the RGB-D camera has its own illumination, we were able to register point clouds even with sparsely illuminated environments. To test the proposed approach, an experiment was performed in a poorly illuminated room. We collected 77 frames of the scene with images ranging from well illuminated to complete lack of light. The final alignment is shown in the teaser (Figure 1), making clear that, even with some regions without illumination, it was possible to align the clouds.

## V. CONCLUSIONS

We have proposed a novel lightweight RGB-D descriptor that efficiently combines intensity and shape information to substantially improve discriminative power enabling enhanced and faster matching process. This approach was compared with other descriptors for images, geometry and with the state-of-the-art approach that combine geometry and intensity. Experimental results showed that our approach outperforms all these techniques, in terms of accuracy, CPU time and matching quality. The experiments have demonstrated also that our approach is robust to register scenes with poor illumination and sparsely textured.

The results presented in this work extends the conclusion of [10] and [29] that the arrangement of intensity and shape information is advantageous not only in perception tasks, but it is useful to improve the quality in registration process. Shape and intensity information enable higher 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. Another important drawback in our methodology is due to RGB-D camera limitations. While laser scanners have field of view (FOV) of about 180 degrees, RGB-D sensors have FOV of 60 degrees. And the maximum distance typically less than 5m for RGB-D. Morevover, the currently RGB-D sensors are confined to indoor scenes.

## ACKNOWLEDGMENT

The authors greatfully acknowledge the financial support of **CNPq**, **CAPES**, **UFMG/PRPq** and **FAPEMIG**. Antônio Wilson Vieira is also affiliated to CCET, Unimontes, MG, Brazil.

## REFERENCES

[1] D. G. Lowe., "Distinctive image features from scale-invariant keypoints," *Internationl Journal of Computer Vision*, pp. 91–110, 2004.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, June 2008.

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

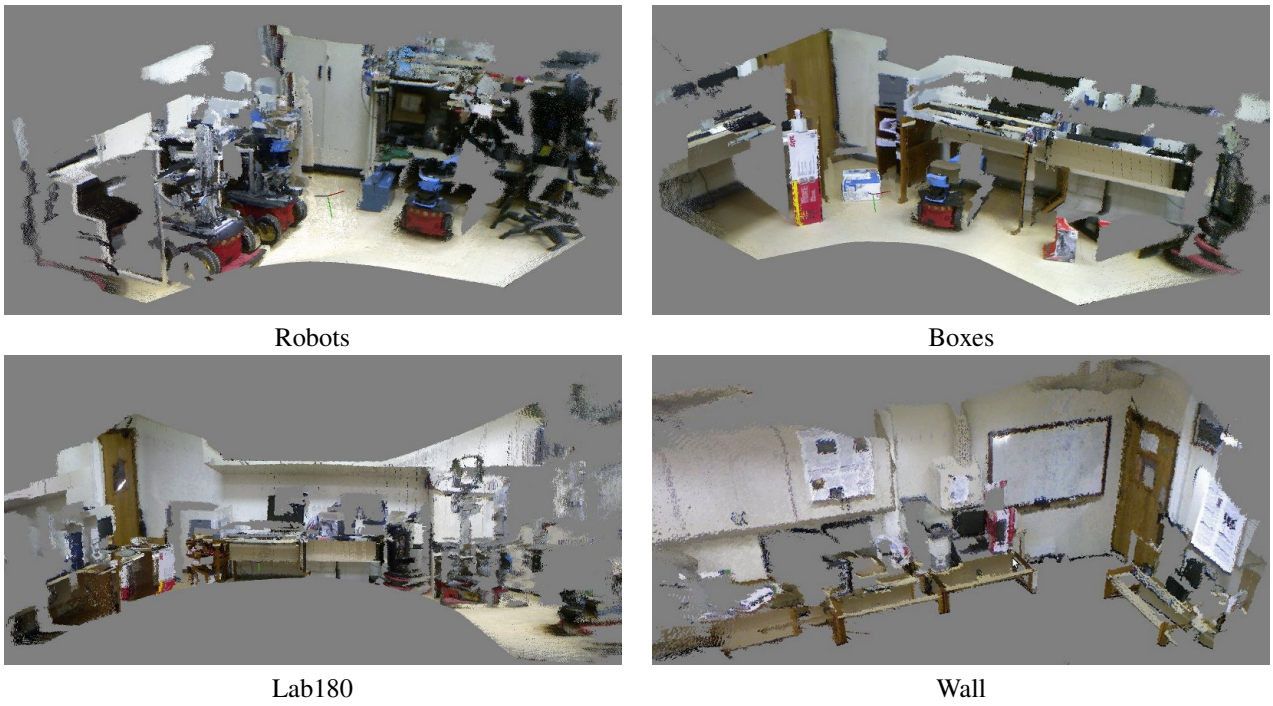



Fig. 10. Data used in the alignment tests. The images show clouds aligned using the BASE descriptor.


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

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

[6] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Trans. PAMI*, vol. 21, no. 5, pp. 433–449, 1999.

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

[8] 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.

[9] 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.

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

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

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

[13] Microsoft, "Microsoft kinect," http://www.xbox.com/en-US/kinect, February 2011.

[14] T. Vieira, A. Peixoto, L. Velho, and T. Lewiner, "An iterative framework for registration with reconstruction," in *Vision, Modeling, and Visualization 2007*. Saarbrcken: Pirrot, november 2007, pp. 101–108.

[15] A. Makadia, E. P. Iv, and K. Daniilidis, "Fully automatic registration of 3d point clouds," in *Proc. CVPR*, 2006, pp. 1297–1304.

[16] S. Winkelbach, S. Molkenstruck, and F. M. Wahl, "Low-cost laser range scanner and fast surface registration approach," in *DAGM Symposium Symposium for Pattern Recognition*, 2006, pp. 718–728.

[17] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. PAMI*, vol. 14, pp. 239–256, February 1992.

[18] Z.-Q. Cheng, W. Jiang, G. Dang, R. R. Martin, J. Li, H. Li, Y. Chen, Y. Wang, B. Li, K. Xu, and S. Jin, "Non-rigid Registration in 3D Implicit Vector Space," in *Proceedings of the 2010 Shape Modeling International Conference*, 2010, pp. 37–46.

[19] A. Sehgal, D. Cernea, and M. Makaveeva, "Real-time scale invariant 3d range point cloud registration," in *Proc. ICIAR*, 2010, pp. I: 220–229.

[20] J. Salvi, C. Matabosch, D. Fofi, and J. Forest, "A review of recent range image registration methods with accuracy evaluation," *Image and Vision Computing*, vol. 25, no. 5, pp. 578 – 596, 2007.

[21] 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.

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

[23] 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.

[24] F. Tombari, S. Salti, and L. D. Stefano, "Unique Signatures of Histograms for Local Surface Description," in *Proc. ECCV*, 2010.

[25] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast keypoint recognition using random ferns," *IEEE Trans. PAMI*, pp. 448–461, 2010.

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

[27] Intel, "SS4 Programming Reference," http://software.intel.com/file/18187, 2007.

[28] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[29] 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 *Proc. ISER*, 2010.

[30] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[31] 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*, Los Angeles, USA, June 2011.

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

[33] K. Mikolajczyk and C. Schmid, "A Performance Evalution of Local Descriptors," *IEEE Trans. PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

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