Variable Dimensional Local Shape Descriptors for Object Recognition in Range Data

Babak Taati[1] Michel Bondy[2] Piotr Jasiobedzki[2] Michael Greenspan[1,3]
[1]Department of Electrical and Computer Engineering
[3]School of Computing
Queen's University, Kingston, Ontario
[2]MDA Space Missions, Brampton, Ontario

Abstract

We propose a new set of highly descriptive local shape descriptors (LSDs) for model-based object recognition and pose determination in input range data. Object recognition is performed in three phases: point matching, where point correspondences are established between range data and the complete model using local shape descriptors; pose recovery, where a computationally robust algorithm generates a rough alignment between the model and its instance in the scene, if such an instance is present; and pose refinement. While previously developed LSDs take a minimalist approach, in that they try to construct low dimensional and compact descriptors, we use high (up to 9) dimensional descriptors as the key to more accurate and robust point correspondence. Our strategy significantly simplifies the computational burden of the pose recovery phase by investing more time in the point matching phase. Experiments with Lidar and dense stereo range data illustrate the effectiveness of the approach by providing a higher percentage of correct matches in the candidate point matches list than a leading minimalist technique. Consequently, the number of RANSAC iterations required for recognition and pose determination is drastically smaller in our approach.

1. Introduction

Model-based object recognition is defined as the 3D registration problem between a range image, known as the *scene* point cloud, and a complete *model*. That is, given two point clouds, one for the model and one for the scene, we are interested in finding the rigid transformation that aligns the model with its instance in the scene data if such an instance exists.

A 3D rigid transformation has 3 positional and 3 rotational degrees of freedom, and exhaustive search through this 6D pose space is infeasible. Efficient pose determination is thus often performed in phases: *point matching*, *pose recovery*, and *pose refinement*. In the point matching phase, correspondences are established between scene points and model points by comparing the local geometrical shapes of various regions of the two data sets. Since the output of first phase often contains a large percentage of incorrect matches (*i.e. outliers*), a statistically robust algorithm such as RANSAC [4] is utilised in the pose recovery phase to find a rough alignment between the model and its instance in the scene. In the final step, a pose refinement algorithm such as the Iterative Closest Point (ICP) algorithm [2] is used to refine the acquired pose.

LSDs are representations that encapsulate local geometries and their similarity is used to hypothesize correspondences between model and scene points. They are computed offline for several feature points on the model and online for some scene feature points. Ideally, LSDs should be highly descriptive so that their similarity is commensurate with the geometric similarity of the local neighbourhoods they represent. They should also be invariant to certain transformations (minimally, rigid transformations) and robust with respect to reasonable levels of noise, changes in the viewing angle, occlusion and self-occlusion, clutter, lighting effects, changes in resolution, and other non-ideal conditions that may arise.

Johnson and Hebert [6] construct their LSDs, known as *Spin Images*, as 2D histograms of distances of neighbouring points from the normal vector and the tangent plane. Chua and Jarvis construct *Point Signature* LSDs as 1D arrays based on the distance profile of the intersection of a sphere with the object from the tangent plane [3]. Other local shape descriptors developed for pose estimation include *Surface Signatures* by Yamany and Farag [15], *Pairwise Geometric Histograms* by Ashbrook *et al*. [1], *Statistical Matrices* by Xiao *et al*. [14], *Point Fingerprints* by Sun *et al*. [10], PCA-Based Descriptors by Taati and Greenspan [13], *Tensor-Based Correspondence* by Mian *et al*. [9], and *Re-*

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

gional Point Descriptors by Frome *et al*. [5]. A review of some of these techniques is offered in [8].

The work presented herein is based on PCA-based descriptors of Taati and Greenspan [13] and offers a generalization that subsumes a large class of local shape descriptors, including many of the aforementioned techniques, such as the Spin Images, Point Signatures, and Pairwise Geometric Histograms. Most descriptors take a minimalist approach, in that they try to construct compact low dimensional descriptors that are fast to compute and compare. In contrast, we take a maximalist approach in LSD generation and focus on enhancing the robustness of the point matching phase, and the time efficiency of the overall pose acquisition scheme, rather than only reducing the computational time of the point matching phase. We note that many of the available descriptors provide a rather low percentage of inliers in their point matching list. For instance, Spin Images, perhaps the most well-known descriptor to date, often provide around $10-15\%$ inliers [6]. As a result, even though establishing the point correspondence list could be performed fast, the RANSAC phase will be relatively expensive, particularly for complex scenes. Our approach is fundamentally different in that we have focused our research away from developing low dimensional and compact descriptors and towards descriptors that could be high dimensional, but that are also highly descriptive and thus can provide a higher percentage of inliers. Consequently, the pose recovery phase could be performed much faster and time efficiency could be achieved in the overall pose acquisition.

Our experiments with simulated and real range images acquired with Lidar and dense stereo have confirmed that our LSDs provide higher percentage of inliers in the candidate point correspondence list than Spin Images and drastically reduce the number of RANSAC iterations.

2. Variable-dimensional descriptors

We generate the LSD for a point based on invariant properties extracted from the principal component space of the local neighbourhood around that point. The 3×3 covariance matrix of local neighbourhood around each point is computed and their Eigenvalue decomposition is used to associate an orthonormal frame $(\vec{i}, \vec{j}, \vec{k})$ and three Eigenvalue scalars (e_1, e_2, e_3), representing vector lengths along each respective axis of the frame, to each point. Using these vectors and scalars, we generate several *properties* for all points and can form a variety of histograms that carry various levels of information about the local geometry.

The orthonormal frame of each point, known as the *Principal Component Space* (PCS) frame, is here treated as a Cartesian coordinate frame. While constructing a descriptor for a point, we refer to the point as the *reference point* p') and to its frame $F(p')$ as the *reference frame*. For each neighbouring point p, the properties define its relationship to the reference point and consist of *position* scalars, *direction* scalars, and *dispersion* scalars. The neighbourhood about point p' is a subset of the entire point cloud and is denoted by $\mathbf{N}_r(p)$, where r is the neighbourhood radius.

Position Property	Description
x, y, z	coordinates along main axes
X_a, Y_a, Z_a	distance to axes $X_a = \sqrt{y^2 + z^2}$ $Y_a = \sqrt{z^2 + x^2}$ $Z_a = \sqrt{x^2 + y^2}$
d_{ist}	distance to reference point $d = \sqrt{x^2 + y^2 + z^2}$

Table 1. Position Properties.

Direction Property	Description
C_θ, C_ϕ, C_ψ	inner products of axes $C_\theta = \vec{i} \cdot \vec{i}'$ $C_\phi = \vec{j} \cdot \vec{j}'$ $C_\psi = \vec{k} \cdot \vec{k}'$
$r_{oll}, p_{itch}, y_{aw}$	ZYX Euler angles
α, β, γ	ZYZ Euler angles

Table 2. Direction Properties.

2.1. Properties

The coordinates of each neighbouring point expressed in the reference frame, x, y, and z along the major $\vec{i}$, semi-major $\vec{j}$, and minor $\vec{k}$ axes, form the three basic position properties. Several other position properties can be calculated based on these coordinates. Tab. 1 lists some possible position properties and the nomenclature we use for referring to them. Note that only three independent position properties exist and therefore there is no benefit in using LSDs that contain more than three position properties.

The rotation that aligns the orthonormal frame of a neighbouring point with that of the reference frame can be defined with three parameters. The inner products of the corresponding axes between the two frames form three direction properties. The rotation can be represented in various forms and therefore it is possible to construct several more properties. Some of these direction properties are listed in Tab. 2, where C_ω is the shorthand notation for $\cos(\omega)$.

Eigenvalues of the neighbourhood covariance matrix form the three basic dispersion scalars. Three scale independent dispersion properties could be generated by normalizing the basic values by their corresponding dispersion property of the reference point. Tab. 3 lists the dispersion properties, where e'_i refers to a basic dispersion property of the reference point.

Dispersion Property	Description
e_1, e_2, e_3	eigenvalues $\lvert e_1\rvert > \lvert e_2\rvert > \lvert e_3\rvert$
$\hat{e}_1, \hat{e}_2, \hat{e}_3$	$\hat{e}_1 = e_1/\acute{e}_1$ $\hat{e}_2 = e_2/\acute{e}_2$ $\hat{e}_3 = e_3/\acute{e}_3$

Table 3. Dispersion Properties.

$\forall\, p \ \in PointCloud$:
$N = \mathbf{N}_r(p) = \{\forall q \in PointCloud \mid ||p - q|| \leq r\}$
$C = covariance(N)$
$EigenDecom(C) \rightarrow F(p) \begin{cases} (\vec{i},\vec{j},\vec{k}) \\ (e_1,e_2,e_1) \end{cases}$

$\forall\, p' \ \in SalientPointCloud$:
$L(p') = \{Properties(F(p'), F(p)) \mid \forall p \in \mathbf{N}_R(p')\}$
$LSD(p') = hist(L(p'))$

Table 4. LSD Generation Pseudo-code.

2.2. Property subset selection

Selecting the property set that is used in generating LSDs for a particular model is of crucial importance to the effectiveness and robustness of point matching. All of the listed properties in Tables 1-3 are robust to lighting conditions, sampling resolution, and other nonideal conditions. However, in the presence of certain conditions such as noise, (self-)occlusion, or clutter, some of them might perform more robustly than others. Furthermore, the descriptive power of these properties are not equal and some might better represent the local geometries. The optimal property subset might not be the same for different models and could be geometry dependant (as is confirmed by the experiments presented in Sec. 3). For example, angular objects could be better described with one set of properties and freeform objects with another. In general, each model could have its own optimal subset. To make the matters more complicated, the optimal subset could also be sensor dependent. This is because different types of non-ideal conditions appear when different sensors are used for image acquisition. For instance, dense stereo range images tend to have less coverage than Lidar data, particularly in texture-less areas. For simplicity, experiments reported in this paper ignore the sensor dependency and assume the same optimal subset for each object for point matching across various data acquisition modes.

The optimal property subset is selected offline for each object. Exhaustive experimentation with all subsets is not applicable since the total number of possibilities is too large (*i.e.* $C_1^{22} + C_2^{22} + \ldots + C_9^{22} > 10^6$). To cope with this issue, we utilised a forward selection scheme [13] that selected the near-optimal subset for each object.

2.3. Salient feature selection

The PCS frame is generated for all the points in the point cloud. However, LSDs need not be constructed for every point. Point clouds in our experiments contained as high as $150,000$ points, while in most cases, a few thousand model LSDs and a few hundred scene LSDs are sufficient for efficient and accurate point matching and pose recovery. *Feature points*, *i.e.* those with an LSD, could be selected at random, or better, based on some efficient prefiltering.

In our implementation, we select salient feature points based on ratios of basic dispersion properties, *i.e.* e_1, e_2, and e_3. For instance, e_3 is zero for a completely planar neighbourhood and is very small for nearly flat areas. As such, by discarding points with a low e_3/e_1 ratio, we avoid generating LSDs for featureless areas and only select the salient subset of the point cloud for LSD generation. The assumption is that unlike points on flat and featureless areas, salient points have a unique geometry and thus lead to more descriptive LSDs. If the number of salient features is still too high (*e.g.* $> 5,000$), then a subsequent downsampling can be applied to reduce the number of feature points to $\sim 2,000$. Since we already compute the dispersion properties for all points, salient feature selection comes at minimal extra computational cost, while it provides a significant improvement over random feature point selection.

2.4. LSD generation

To generate an LSD for feature point p', first a local neighbourhood with radius R is found around p'. (Note that R is not necessarily the same as r.) Properties of all points in this neighbourhood can be extracted with respect to p'. A subset of these properties are then used to generate the LSD for p'. Since only nine independent properties exist for each point we limit the maximum dimensionality of our LSDs to nine. The desired set of $d \in [0, 9]$ properties are determined offline for each object and only those properties are extracted in the online computation. A d-dimensional histogram of these properties constitutes the LSD for point p'. A formal description of the aforementioned LSD generation scheme is presented as pseudo-code in Tab. 4.

2.5. Histogram representation of LSDs

The final stage is to build a scalar-quantized or vector-quantized histogram based on this list. For scalar quantization, each axis along the d-dimensional space of the property list is divided into several segment and thus the space is partitioned into several bins in the shape of hyper-cubes. Vector quantization provides an alternative histogramming technique that avoids the exponential rise in the number of bins with the dimensionality of the property space. Property lists of all model feature points are computed offline and merged together to form a large list. A cluster finding algorithm is used to find several cluster centres in the

merged list and the space is partitioned about these cluster centres. The contribution of each neighbour point to the vector quantized histogram is then computed by incrementing the bin corresponding to the cluster centre that is closest the d-dimensional property space. To ease the online burden, a *k*d-tree could be computed offline based on the cluster centres so that the nearest neighbour computation could be performed efficiently, nearly as fast as binning for scalar-quantized histograms. We performed our experiments with both scalar and vector quantization, using several code book lengths.

2.6. Similarity Measures

Two metrics were used to measure the degree of similarity between two LSDs. The first was histogram intersection (I), which is a measure of the number of bins that both histograms have in common and provides a continuous similarity measure in the $[0, 1]$ interval, where 1 denotes identical histograms [11].

The similarity of all scene LSDs to all model LSDs could be measured using the I metric to generate an $n_M \times n_S$ confusion matrix, where n_M and n_S represent the number of model and scene descriptors respectively.

The second metric is an ambiguity measure based upon the ratio of the top two ranking histogram intersections. It was recognized by Lowe [7] that some LSDs may be very similar to a number of other LSDs. When this occurs, it is more likely that the matching will be incorrect, even if the similarity measure I is high. The ratio of the first and second highest ranking LSDs is therefore considered as a measure of ambiguity. If this ratio is high, then there is a good likelihood that a correct match has been determined; otherwise, the match is ambiguous and should be discarded. To formalize, let I_1 and I_2 be the value of the first and second top matching model LSDs for a particular scene LSD, using the I measure. The ambiguity measure S is normalized so that its range is from zero to one, *i.e.* $S = (I_1 - I_2)/I_1$.

An absolute threshold I_t or S_t could be set, above which two LSDs are deemed to be similar and form a possible match between their corresponding points.

2.7. Subsumption of minimalist LSDs

The described LSD generation scheme based on property lists can lead to a large variety of descriptors and offers a generalization that subsumes the majority of existing descriptors developed by various researchers. For example, Spin Images [6] are 2D histograms of distances of neighbour points to the surface tangent plane and the surface normal. In our notation, these translate to the z and Z_a properties and thus Spin Images are in essence 2D LSDs based on the [z , Z_a] property subset. Likewise, Pairwise Geometric Histograms [1] generate descriptors based on the [Z , θ] property subset. Point Signatures [3] are constructed with distances from the tangent plane along the intersection of a sphere of radius R with the surface (*i.e.* $d_{ist} = R$), albeit not in the form of a histogram, and are therefore based on the $[z, d_{ist}]$ property subset. Tensor-Based descriptors [9] are computed from surface meshes rather than point clouds, but they are very similar to 3D histograms based on $[x', y', z]$ properties, where x' and y' are position properties computed from rotating x and y properties about the k axis.

(a) (b) (c)

Figure 1. Sample Lidar range images.

3. Experiments

The primary motive of our research has been developing a reliable and efficient pose acquisition technique for localization and tracking of geostationary satellites for the purpose of automatic rendezvous and docking by unmanned spacecrafts. For this reason, we have performed extensive experiments on Lidar and dense stereo images of a satellite model . We had access to a complete CAD model and also a physical mock-up for Lidar and stereo image acquisition. The mock-up was mounted on a robotic arm that was commanded to certain configurations for image capture. Ground truth location and orientation of the mock-up was therefore roughly known which was later refined using ICP refinement. Fig. 1(a) illustrates a sample Lidar image of the satellite model. To further test the performance of our method on other models, we have used the object recognition data set available at the website of University of Western Australia (UWA) [9] and report sample detection and localization results in cluttered scenes and in presence of partial occlusion by other objects. The data set contains five models and fifty Lidar images of these objects in various placements. Fig. 1(b) illustrates a sample scene range image where the five models are present: chef (top left), chicken (bottom left), two dinosaurs (top right), and a rhinoceros (bottom right). For brevity, only sample results from localizing the chef and chicken models are presented. Ground truth alignments were generated manually in cases for which they were not available.

Dimensions of the satellite mock-up was in the order of $2m$. Dimensions of the other models were in the or-

der of $20cm$. Scalar quantization was tested on all the objects. Satellite experiments were also performed with vector quantized representation of the LSD space. All experiments were repeated for our implementation of Spin Images, enhanced with rejecting ambiguous point matches using the S similarity measure. Code book sizes of 32, 64, 128, 256, 512, 1024, and 4096 were tried for vector quantization. For brevity, only results with code book size 256 are reported since vector quantized histograms with 256 bins often performed well compare to other code book sizes. Spin Images were represented in these experiments by 16×16 histogram, which made each Spin Image the same size in memory as the vector quantized $6D$ LSDs, *i.e.* a vector of length 256. We emphasize that this level of binning avoided too fine binning that Spin Images are sometimes prone to suffering from.

The code was implemented in C/C++ using the OpenGL and Coin3D libraries for visualization. Implementation was efficient but without use of any hardware acceleration. The entire online pose acquisition process runs at $\approx 0.2-0.5Hz$ for scenes containing $\approx 50,000$ $3D$ points, including background clutter. The online process includes stereo image acquisition, dense stereo processing, scene LSD generation and comparison with model LSDs for point matching, pose generation, and pose refinement.

A forward selection scheme [13] was utilised to select a property subset for each object. For the satellite model, feature selection was performed based on manually selected points on a single range image and their ground truth match on the model. For the other objects, two range images of each object from slightly different view point were used for property selection. Details of the selection process could be found in [13, 12]. In all the experiments points that were matched within a small distance d_t of their ground truth match were considered inliers. The d_t threshold was set experimentally to $5cm$ for the satellite at $5mm$ for the other objects. A 6D LSD based on the $[x, X_a, Y_a, e_1, e_2, e_3]$ properties was selected for the satellite according to this scheme. 4D and 3D LSDs based on $[Z_a, d_{ist}, r_3, e_1]$ and $[Z_a, d_{ist}, r_3]$ properties were selected for the chef and chicken models respectively. These LSDs were used throughout the remaining experiments.

For scalar quantization, the number of bins along each dimension were set to 4, 8, and 16 for the satellite, chef, and chicken respectively. r and R radii were respectively set to $15cm$ and $30cm$ for the satellite model and to $2cm$ and $5mm$ for the other models. Fig. 2 illustrates sample results from experiments that lead to the selection of these radii values. It shows the results of a sample point matching test between the satellite model and a Lidar image. In this case, 2,040 and 284 LSDs were generated on the model and the scene data respectively. The top 100 matches were selected according to the I similarity measure. The figure shows a

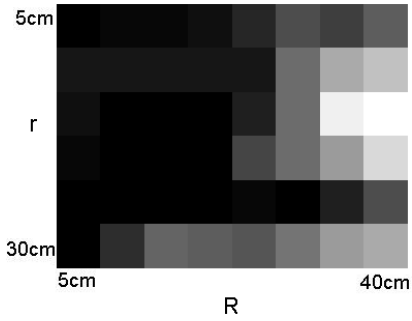

Figure 2. Changes in the number of inliers with respect to changes in r and R values.

larger number of inilers with lighter colours and is scaled in the intensity range for better viewing. The radii r and R were tested in the range of $[5cm, 30cm]$ and $[5cm, 40cm]$ respectively in increments of $5cm$ to form a matrix of 48 elements. It can be observed that the largest (*i.e.* brightest) matches occur on the third row ($r = 15cm$) and the eighth column ($R = 40cm$) in this particular case. Here, the lightest matrix cell corresponds with 33 correctly identified matches.

It was observed that in most cases, increasing the value of r initially increased and then decreased percentage of inliers, leading to a peak value such as the one at $15cm$ in the above example. The effect of increasing the R radius on the other hand was monotonous as it often increased the point matching precision in the absence of clutter from other objects in the scene. This could be explained by the fact that changing r affects the points that are selected for computing the principal component space and by selecting a too large value, self-occlusion starts to affect the accuracy of the reference frames. Changing R on the other hand affects the points that contribute to the LSD of each reference point and therefore increasing it has no negative effect as long as there is no clutter in the scene. Since the Lidar image that was used to generate the above figure did not include any clutter it is natural that increasing the radius only improved point matching. We also note that increasing the radii affects the computational cost of the algorithm and it is desirable to avoid very large values.

Since number of RANSAC iterations is highly dependant on the inlier percentage, we used the ability of the system to find the correct pose within a certain number of iterations as our first performance measure. However, since RANSAC involves a random phase, we also used *precision* and *recall* for evaluating the performance of point matching as they are the standard metrics in matching tasks. Precision is the ratio of the number of correct matches (*i.e.* inliers), over the total matches retrieved. Recall is similar to precision, except that it considers the total number of possible correct matches. A standard way to visualize these measures is to plot recall

against (1-precision).

In the ideal case, the data would fall in the top left corner, *i.e*. perfect recall, and perfectly precise. In general, the closer that the curve is to the top left corner, the more desirable the operating point. In $3D$ point matching however, recall is typically low and the percentage of inliers (*i.e*. precision) is often around $5-15\%$ [6] causing the plots to often fall closer to bottom right corner. We are more interested in precision than recall, because only a small number (*i.e*. 3) of correct matches can lead to a successful RANSAC-based pose generation. In other words, if recall were small so that only 3 matches were retrieved, this would be entirely suitable, and indeed even preferable, so long as those matches were correct. The independent variable for the curves is the threshold on the similarity measure I_t or S_t.

Our first set of experiments included image to image registration for the satellite model and was aimed at measuring the effectiveness of LSDs for tracking applications. We note that similar image to image registration could also be used in terrestrial applications such as model building. Five dense stereo data sets, each comprising of two point clouds, were acquired of the satellite model. The range image pairs from each set were acquired from slightly different viewpoints.

The five cases viewpoint changes included a $10cm$ shift along the depth axis (*Small Depth Translation*), $50cm$ shift along the lateral axis (*Large Lateral Translation*), 10° and 30° rotations about the yaw axis (*Small* and *Large Rotation*), and no shift between the two view points (*Same Pose*). The Same Pose case set was used to establish a performance limit, as this was the scenario where the matching was expected to be the most robust.

A set of 100 feature points were randomly selected from the satellite CAD model. Each range image was then aligned with this CAD model, and the 100 points in each image that corresponded to the manually selected feature points were identified. The 6D LSDs were generated for each feature point in both range images, and the LSDs of the two images were compared to form a 100 confusion matrix. The best matches were then selected according to both I and S similarity measures. Figs. 3 and 4 illustrate sample results from these test.

It could be observed that according to both the I and the S measures, vector quantization performs nearly as well as scalar quantization and they both perform better than the Spin Images, *i.e*. their recall vs. (1-precision) curves are further to the left and to the top. Note that using the S_t threshold, the number of returned point matches is in the range of $[0, n_S]$, where n_S is the number of descriptors on the image that is treated as the scene point cloud. As such, the recall values do not necessarily reach the 100% level (and the curves do not reach the top right corner) since even at the lowest levels of the threshold only a limitted number

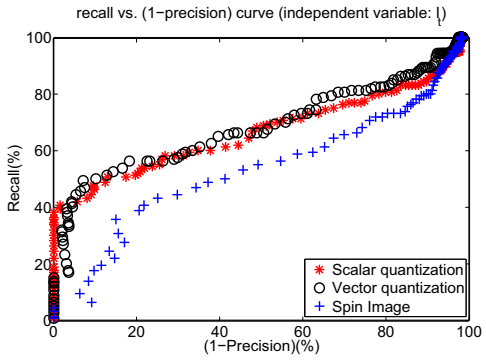

Figure 3. *Large Latteral Translation* between two dense stereo images.

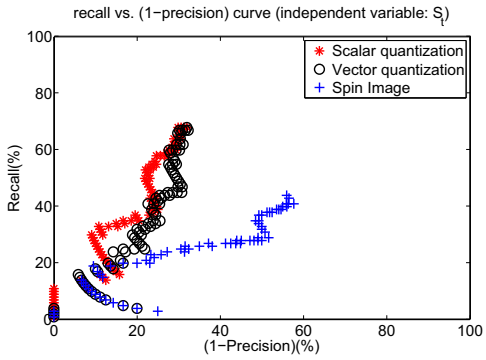

Figure 4. *Small Rotation* between two dense stereo images.

of matches are reported.

Two of the above test cases, the Small Depth Translation and the Large Rotation, were repeated using Lidar data. A sample result from these test are shown in Fig. 5. As this figure illustrates, the recall levels are significantly lower than the dense stereo case. This is rather counter intuitive since one expects Lidar point clouds to be of better quality that dense stereo images. However, the acquired Lidar images in this case had rather poor quality in that it contained very few points on the solar panels of the satellite model and on parts of its main body. This was perhaps due to large changes in reflexive properties of the surface of mock-up. The main body for instance, was covered with light coloured heat insulating sheets that were highly reflective while the solar panels were dark and less reflective. This highly non-uniform sampling of the surface , as seen in Fig. 1(a), lead to the rather lower than expected performance of both the LSDs and the Spin Images. But it could still be observed that the LSD point matching outperformed the Spin Images.

For the second part of our experiments we performed LSD-based point matching between model point clouds and both Lidar and dense stereo range images. For these experiments, the number of model and scene descriptors were

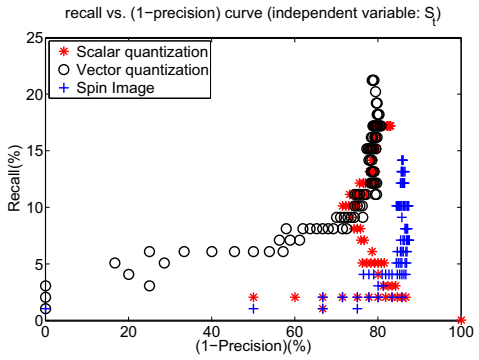

Figure 5. *Large Rotation* between two Lidar images.

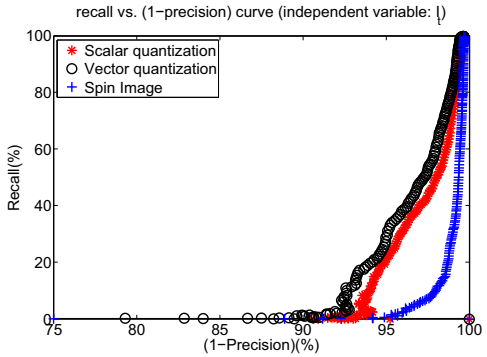

Figure 6. Recall vs. (1-precision) curve for point matching between a Lidar image and the complete satellite model.

about $2,000$ and $200-700$ respectively. This lead to a dense coverage of model salient feature, which were processed offline, and a rather sparse set of feature points on the scene. For the satellite, the range data consisted of about $30,000$ points for the model point cloud and $10,000$ to $25,000$ points for the scene image, depending on the range image acquisition mode. For the UWA objects, the models contained about $150,000$ points each and Lidar range images contained $100,000$ to $150,000$ points.

Fig. 6 shows sample satellite pose acquisition results comparing the performance of our 6D LSD with the Spin Images and illustrate that the Spin Images are again outperformed. It could also be observed that the precision levels are lower than the image to image registration cases as the image to model case is more complicated. This is because in image to image point matching, the point clouds are obtained from the same mode of acquisition (*e.g.* Lidar) and they often share a the same levels of noise, clutter, etc., even when they are taken from viewpoints with large disparity.

Fig. 7 illustrates sample object recognition results for the models from the UWA data set. The precision plots compare the performance of our LSDs with Spin Images for the chef (top) and the chicken (bottom) models in the Lidar im-

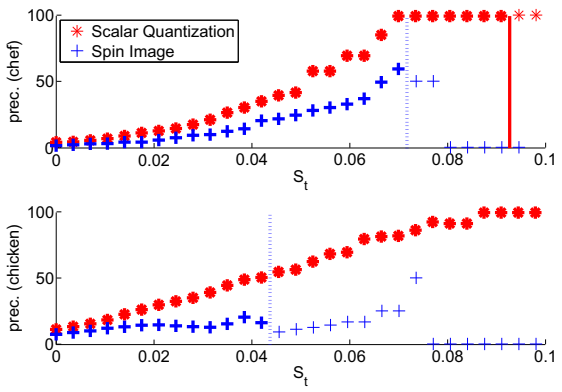

Figure 7. Precision for chef (top) and chicken (bottom).

age shown in Fig 1(b). In this image, clutter and occlusion values, as defined in [6], are approximately 74% and 79% for the chef and 71% and 52% for the chicken. Precision values for which the number of correct returned matches are enough for RANSAC (*i.e.* ≥ 3) are shown with bold markers. A vertical line indicates the first S_t value for which the number of correct returned matches drops below 3 for Spin Images (dotted) and for LSDs (solid). As the figure illustrates, the LSD precision values are larger than that of Spin Images. More importantly, for LSDs, increasing the S_t threshold is significantly more effective in rejecting outliers matches while maintaining the minimum level of required correct matches. In both cases, LSD precision values reach 100% while for Spin Images the number of correct matches drops below 3 much earlier.

Tab. 5 presents the average (*avg*) and standard deviation (*stdev*) of number of RANSAC iterations over 20 trials for detecting and locating the chef and the chicken in the same image for two different values of S_t using both LSDs and Spin Images (*SI*). As expected, the numbers are significantly smaller when LSDs are used. RANSAC stopped as soon as the object was found (*i.e.* a significant number of model points were aligned to the scene within a small distance threshold) or if the object was not found after $1,000$ iterations. The Table also shows that the number of failures, *i.e.* cases were the object was not found, were drastically smaller for LSDs. 3D points corresponding to the chef and chicken are segmented once they are identified. Fig. 1(c) shows the chef and chicken models overlayed on the range data to illustrate that the object recognition task is done correctly.

The following observations were made from our experiments. In the majority of the cases, both vector quantized and scalar quantized LSDs outperformed the Spin Images resulting in higher precision levels and smaller number of RANSAC iterations. Furthermore, the vector quantized

	S_t	0.04		0.06	
	Descriptor	LSD	SI	LSD	SI
Chef	*avg*	24.8	196.9	3.4	41.1
	stdev	25.3	230.0	2.1	33.7
	# fail.	2	9	0	0
Chicken	*avg*	11.2	*fail*	3.6	74.3
	stdev	9.0	*fail*	3.5	68.7
	# fail.	0	20	0	3

Table 5. *Avg* and *stdev* of number of RANSAC iterations and number of failures over 20 trials.

LSDs performed nearly as well, an in fact in some cases even better, than the scalar quantized LSDs. Therefore, they are preferable to the scalar quantized version as they provided the same performance at a lower memory and computational cost. The S similarity measure, based on rejecting ambiguous matches, was nearly always better than the I similarity measure which was based on simple histogram comparison. As we increased the S_t threshold, precision often improved, at least up to a certain point and then either reached 1 or dropped to zero. It was also observed that increasing the S_t threshold for enhancing precision was more effective for LSDs than it was for Spin Images. Too high values of S_t translate to rejecting all matches as ambiguous and returning no matches where precision is undefined. Values of S_t close to this level render the precision unreliable, even if it is close to 100% since the total number of returned matches are very small (*e.g.* ≤ 5). Therefore, it is best to set the threshold at a moderate value (around $0.06 - 0.08$) that leads to high levels of precision along with high number of returned matches.

4. Conclusions and future works

We presented a generalization scheme for point matching with local shape descriptors that subsumes the majority of previously developed descriptors and provides a framework for a systematic comparative study. We also introduced the concept of maximalist descriptors, where descriptiveness and robustness of LSDs are emphasized rather than their compactness. We tune our LSDs for each model by selecting their near-optimal property subset. This involves pre-computation on each model which is aimed at easing the online computation performed on the scene data.

While more descriptive LSDs could take longer to generate or compare and as such increase the required time for point matching, it was experimentally confirmed that they enhance the percentage of inliers in the point correspondence list and thus significantly reduce the number of RANSAC iterations in the pose recovery phase. For future works, we intend to compare the performance of our LSDs with Regional Point Descriptors [5] and Tensor-Based Descriptors [9].

References

[1] A. Ashbrook, R. B. Fisher, C. Robertson, and N. Werghi. Finding surface correspondence for object recognition and registration using pairwise geometric histograms. *Proc. 5th European Conferene on Computer Vision*, pages 674–680, 1998. 1, 4

[2] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence, 14(2)*, pages 239–256, 1992. 1

[3] C. S. Chua and R. Jarvis. Point signatures: a new representation for 3d object recognition. *Intl. Journal of Computer Vision, 25(1)*, pages 63–85, 1997. 1, 4

[4] 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*, pages 381–395, 1981. 1

[5] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. Recognizing objects in range data using regional point descriptors. *Proc. European Conf. Computer Vision, Prague, Czech Republic*, 2004. 2, 8

[6] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence, 21(3)*, pages 433–449, 1999. 1, 2, 4, 6, 7

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision, 60(2)*, pages 91–110, 2004. 4

[8] A. S. Mian, M. Bennamoun, and R. Owens. Automatic correspondence for 3d modeling: An extensive review. *Intl. Journal of Shape Modeling, 11(2)*, pages 253–291, 2005. 2

[9] A. S. Mian, M. Bennamoun, and R. Owens. 3d model-based object recognition and segmentation in cluttered scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence, 28(10)*, pages 1584–1600, 2006. 1, 4, 8

[10] Y. Sun, J. A. Koschan, D. L. Page, and M. A. Abidi. Point fingerprints: A new 3-d object representation scheme. *IEEE Trans. Systems, Man, and Cybernetics, Part B, 33(4)*, pages 712–717, 2003. 1

[11] M. J. Swain and D. H. Ballard. Color indexing. *Intl. Journal of Computer Vision 7(1)*, pages 11–13, 1991. 4

[12] B. Taati, M. Bondy, P. Jasiobedzki, and M. Greenspan. Automatic registration for model building using variable dimensional local shape descriptors. *Proc. 6th Intl. Conf. 3D Digital Imaging and Modeling*, 2007. 5

[13] B. Taati and M. Greenspan. Satellite pose acquisition and tracking with variable dimensional local shape descriptors. *Proc. IEEE/RSJ IROS 2005, Workshop Robot Vision for Space Applications*, pages 4–9, 2005. 1, 2, 3, 5

[14] G. Xiao, S. H. Ong, and K. W. C. Foong. Three-dimensional registration and recognition method for tooth crown. *Proc. Intl. Congress on Biological and Medical Engineering - The Bio-Era: New Frontiers, New Challenges*, 2002. 1

[15] S. M. Yamany and A. A. Farag. Surface signatures: An orientation independent free-form surface representation scheme for the purpose of objects registration and matching. *IEEE Trans. Pattern Analysis and Machine Intelligence, 24(8)*, pages 1105–1120, 2002. 1