Thrift: Local 3D Structure Recognition

Alex Flint, Anthony Dick, Anton van den Hengel
School of Computer Science
The University of Adelaide
North Terrace, Adelaide, 5005
South Australia, Australia
{alex.flint,anthony.dick,anton.vandenhengel}@adelaide.edu.au

Abstract

This paper presents a method for describing and recognising local structure in 3D images. The method extends proven techniques for 2D object recognition in images. In particular, we propose a 3D interest point detector that is based on SURF, and a 3D descriptor that extends SIFT. The method is applied to the problem of detecting repeated structure in range images, and promising results are reported.

1. Introduction

Image based object recognition is a long standing central problem in computer vision. Recently, attention has turned to the use of local feature descriptors that, given a keypoint in an image, calculate a signature describing the image about that point. Using local descriptors to describe an object provides robustness to partial occlusion, and depending on the design of the descriptor can provide robustness to changes in illumination and viewpoint.

For example, the Scale Invariant Feature Transform (SIFT) [4] has proven to be a very effective descriptor for object recognition from images. SIFT calculates a signature that characterises the image in the neighbourhood of a keypoint in a way that is robust to changes in global illumination, object rotation and scale. The signature is based on histograms of image grey-level gradients which are calculated at several scales, and normalised with respect to a locally dominant orientation.

The idea of this work is to build a local 3D feature descriptor with comparable robustness to missing data and changes in viewpoint. This was initially motivated by previous work of the authors [5] in image based modelling. In this domain it is common to have a 3D data set—whether captured from a range finder, or the output of structure and motion estimation, or modelled manually—that is incomplete. Often it is the case that this 3D data will contain repeated structure, some instances of which are captured or modelled with higher fidelity than others. If such repetition can be recognised automatically, information from instances that are well modelled can be propagated to those that are poorly modelled, resulting in a more accurate overall model.

Local 3D feature descriptors have been investigated previously. For example, Frome et al. [2] define a *shape context* for a 3D keypoint. The shape context is computed by counting the number of 3D points lying in a neighbourhood of the keypoint. These counts are partitioned into a histogram based on their distance and direction from the keypoint.

In a similar vein, *spin images* [3] also divide the area around a keypoint into a number of spatial bins and then count the number of points in each bin. The difference is that in spin images the bins are defined by height and radius; i.e. the neighbourhood is cylindrical, which has advantages over a spherical neighbourhood. Neither transform is invariant to scale, and, although they exhibit some robustness to rotation (histograms are calculated relative to estimated surface normal) they are sensitive to small changes in the computed surface normal.

The success of local feature descriptors depends strongly on the choice of keypoint locations. In 2D images, good keypoints are those that can be well localised, such as corner points where the intensity gradient is high in all directions. Several techniques such as the Harris corner detector, and more recently SURF [1], have been developed to identify these points. In 3D images, we also require keypoints that can be well localised, but in 3D this requires that the spatial gradient of the surface about a keypoint be high in all three directions.

This paper proposes a new 3D feature descriptor, called ThrIFT, that extends the successful SIFT and SURF algorithms to keypoint selection, identification and matching in range data. It brings many of the advantages of these algo-

0-7695-3067-2/07 $25.00 © 2007 IEEE
DOI 10.1109/DICTA.2007.68

rithms to bear on the problem of 3D structure recognition. We show how this 3D keypoint detector and descriptor can be combined to detect repeated 3D structure in range data of building facades.

The remainder of this paper is organised as follows. Section 2 describes our interest point detector. Section 3 describes our 3D descriptor. In section 4 we present empirical results, and section 5 concludes the paper.

2. A 3D Interest Point Detector

The objective of an interest point detector is to repeatably identify the same scene points under a range of image transformations, such as a change in viewpoint or illumination. This requires that the interest points be located at scene features that define an unambiguous location (corners have this property but edges do not).

SIFT and SURF use the determinant of the Hessian to measure the distinctiveness of candidate interest points. This is successful because when the determinant of the Hessian is large then both principal curvatures are large [4], meaning that the point is well defined and likely to be repeatably detected under different viewing or lighting conditions. Conversely, when the determinant of the Hessian is small then at least one of the principal curvatures is small, so localisation in this direction may vary under image transformations.

In range data, interest points must be well localised in all three dimensions if they are to be repeatably detected at the same location. ThrIFT uses the 3D version of the Hessian to select such interest points. We approximate a density function $f(x, y, z)$ by sampling regularly in space throughout the data (explained in detail below). We then construct a scale space over the density function, and search for local maxima of the Hessian determinant.

2.1. The Density Map

In this work we consider range data to be a set of 3D points:

$$\mathcal{X} = \{x_i \in \mathbb{R}^3\}$$

We wish to approximate a density function $f(x, y, z)$ from this data. Let $n(B)$ be the number of data points in the region $B \subseteq \mathbb{R}^3$. We can approximate f in any such region using

$$\int_B f(\mathbf{x})\, d\mathbf{x} = n(B)$$

We define equal-sized boxes $\mathcal{B} = \{B_{ijk}\}_{(i,j,k)\in I \subset \mathbb{Z}^3}$ and space them regularly in each spatial dimension:

$$\begin{aligned} B_{ijk} = \{(x, y, z) \in \mathbb{R}^3 \mid\ & i\alpha \le x < (i+1)\alpha, \\ & j\beta \le y < (j+1)\beta, \\ & k\gamma \le z < (k+1)\gamma\} \end{aligned}$$

We then construct f as a sum of delta functions:

$$f(x, y, z) = \sum_{(i,j,k)\in I} D(i, j, k)\delta(x - X_{ijk}, y - Y_{ijk}, z - Z_{ijk})$$

where $(X_{ijk}, Y_{ijk}, Z_{ijk}) = \overline{B}_{ijk}$ is centre of the box B_{ijk} and

$$D(i, j, k) = \frac{n(B_{ijk})}{\underset{(i,j,k)\in I}{\operatorname{argmax}} \Big\{ n(B_{ijk}) \Big\}}$$

is the normalised density map. In practice we operate directly on D since it is readily represented as a 3D array. D can be thought of as the 3D analogy to a 2D image: each element represents the density (resp. pixel intensity) in a region of space. We can apply 3D convolutions to D in much the same way as for 2D images.

2.2. Density Scale Space

2D detectors often construct a scale space to enable feature detection at a range of scales. This is achieved by convolving the image with Gaussian kernels of increasing radius, resulting in an image pyramid. We apply a similar concept to search the density map D over a range of scales. We convolve D with a series of 3D Gaussian kernels to construct a pyramid of density maps, with each layer representing the scale $\sigma = k\sigma'$ where σ' is the scale of the layer immediately below. For efficiency we downsample the density map by a factor of 2 when the scale reaches 2 (and simultaneously reduce the variance of the Gaussian kernel by a factor of 2). This bounds the size of the convolution kernels and hence leads to a large performance improvement.

Let $L(x, y, z; \sigma)$ be a scale space for D:

$$L(x, y, z; \sigma) = (D \otimes g(\sigma))(x, y, z)$$

where $g(\sigma)$ is a 3D Gaussian with variance σ:

$$g(x, y, z; \sigma) = \exp\Big(\frac{-x^2 - y^2 - z^2}{2\sigma^2}\Big)$$

(Note that we have omitted the normalisation constant.)

The number of downsampling operations (i.e. the number of octaves), and the number of scales we generate between downsampling (i.e. the number of layers per octave) are user-specified parameters.

2.3. Selecting Interest Points

Interest points must be well localised in all three spatial dimensions in order to be repeatable. We implement this in ThrIFT by choosing points for which all three principal curvatures are large. Such points will represent significant extrema in the density function along all three directions, which will lead to the interest point being detected in

the same position when the scene is viewed under different viewpoint or lighting conditions.

We use the determinant of the 3×3 Hessian matrix to find such points because it can be computed efficiently and accurately, and is defined for arbitrary scale. Given a point $\mathbf{x} = (x, y, z)$ and a scale σ, the Hessian at $(\mathbf{x}, \sigma)$ is defined as:

$$\mathcal{H}(\mathbf{x}, \sigma) = \begin{pmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) & L_{xz}(\mathbf{x}, \sigma) \\ L_{yx}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) & L_{yz}(\mathbf{x}, \sigma) \\ L_{zx}(\mathbf{x}, \sigma) & L_{zy}(\mathbf{x}, \sigma) & L_{zz}(\mathbf{x}, \sigma) \end{pmatrix}$$

where

$$L_{xx}(\mathbf{x}, \sigma) = \frac{\partial^2}{\partial x^2} L(\mathbf{x}, \sigma)$$

is the second partial derivative of L in the x direction, and similarly for the other partial derivatives. In practice we compute the terms in the Hessian by direct convolution of D with Gaussian second partial derivatives:

$$L_{xx} = D \otimes \frac{\partial^2}{\partial x^2} g(\sigma)$$

and similarly for the other partial derivatives. This avoids constructing the scale space and computing the terms in the Hessian as separate operations.

In our implementation we chose to compute exact convolutions using separable kernels (the alternative would be to approximate the kernels with box filters as in SURF). In practice we found that we could use relatively small density maps without loss of performance, and hence the extra efficiency of box filters was unnecessary. Furthermore, using exact convolutions allows us to sample at any desired frequency in the scale domain, which makes it unnecessary to interpolate the location of interest points later in the detection process.

We compute $|\operatorname{Det}(\mathcal{H})|$ at each point in the scale space. To eliminate weak responses we apply a constant threshold T. Next we apply non maximal suppression within a $3 \times 3 \times 3 \times 3$ window. The remaining responses are local maxima of $|\operatorname{Det}(\mathcal{H})|$, and these are exported as interest points.

In SIFT responses with two principal curvatures of different sign are eliminated, since such points represent saddles in the intensity function. We allow such points because in the 3D setting saddles may represent useful interest points.

We could now expand $|\operatorname{Det}(\mathcal{H})|$ about each interest point using the Taylor series, which would allow us to further localise the interest points, as in SIFT and SURF. However, in our approach we sample regularly in the spatial and scale domains, and we found in practice that further localisation was unnecessary.

The ThrIFT detector is summarised by:

$$Interest(\mathcal{X}) = \operatorname*{arglocalmax}_{(\mathbf{x}, \sigma)} |\operatorname{Det}(\mathcal{H}(\mathbf{x}, \sigma))|$$

3. The 3D descriptor

The success of the descriptors used in SIFT and SURF have been partially attributed to the use of image gradients as the basis for describing image patches. Image gradients capture the dominant orientation of blocks of pixels, and are robust to changes in viewpoint and illumination.

ThrIFT also uses orientation information as the basis for its descriptor. In the case of range data, the dominant orientation at a point is the direction of the surface normal at that point. Since we do not have explicit surface normal information we approximate it by fitting a least-squares plane to the points in a sphere centred at the point.

We may think of the surface normal at a point as the principal direction of the density map at that point. In this sense, the surface normal is a direct generalisation of the gradient orientation used in SIFT. Furthermore, the three components of the surface normal vector correspond to a generalisation of the dx and dy image gradients used in the SURF descriptor.

There is a further advantage to using surface normal information that is specific to range data. In real range data, the density of points on a surface is determined by the viewpoint of the camera or range finder. A surface close to the viewpoint will by sampled more densely than the same surface further from the viewpoint, and similarly a surface that is normal to the viewpoint will be sampled more densely than the same surface oblique to the viewpoint (see Figure 1). In fact, as an object rotates, the relative sampling density of its surfaces will change significantly, as each surface changes its orientation with respect to the viewpoint. Hence it is important that our descriptor be robust to such changes in sampling density.

In general the surface normal at a point is unaffected the sampling density at that point. In practice we approximate the surface normal with a least-squares plane, so changes in sampling density will invariably have some effect on the surface normal we compute. However, in the presence of a significant number of points we can expect these errors to cancel out, since all the points are situated on the same underlying surface, and we assume that sensor noise is independent for each data point. This means that the normal to the least-squares plane will be largely unaffected by changes in sampling density. Hence by using surface normals our descriptor becomes more robust to changes in sampling density than other descriptors that use only location information (e.g. spin images [3] and shape contexts [2]).

Our descriptor operates as follows. For each interest point $\mathbf{z} = (\mathbf{x}, \sigma)$ we define the support set:

$$Support(\mathbf{z}) = \{\mathbf{y} \in \mathcal{X} : \|\mathbf{y} - \mathbf{x}\| \leq \sigma\}$$

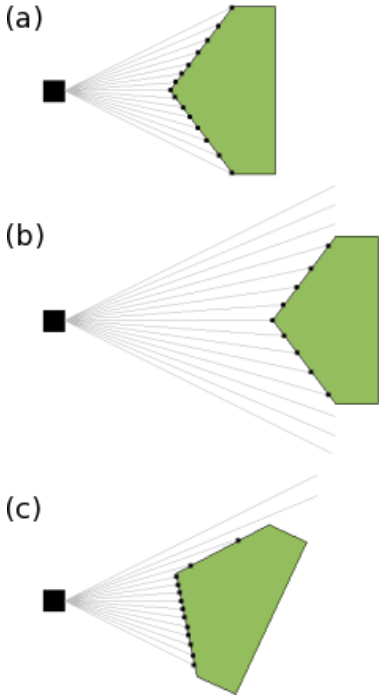

Figure 1. Changes in sampling density resulting from changes in viewpoint

For each $\mathbf{y} \in Support(\mathbf{z})$ we define two windows:

$$W_1 = \{\mathbf{p} \in \mathcal{X} : \|\mathbf{p} - \mathbf{y}\| \leq \omega_{small}\}$$
$$W_2 = \{\mathbf{p} \in \mathcal{X} : \|\mathbf{p} - \mathbf{y}\| \leq \omega_{large}\}$$

Let P_1 and P_2 be the least-squares plane for W_1 and W_2 respectively, and let $\mathbf{n}_{small}$ and $\mathbf{n}_{large}$ be normal to P_1 and P_2 respectively. These two vectors can be interpreted as the principal curvatures of the density map at $\mathbf{y}$ for scale ω_{small} and ω_{large} respectively. In our implementation ω_{small} and ω_{large} are user-defined parameters, but they could also be determined automatically from the detected scale σ. See Figure 2 for a geometric interpretation of these entities.

The descriptor output for the interest point $\mathbf{z}$ is a histogram over the angle θ between $\mathbf{n}_{small}$ and $\mathbf{n}_{large}$ for each $\mathbf{y} \in Support(\mathbf{z})$

$$\cos(\theta) = \frac{\mathbf{n}_{small} \cdot \mathbf{n}_{large}}{\|\mathbf{n}_{small}\|\|\mathbf{n}_{large}\|}$$

The number of bins nb is a user-defined parameter. The bins are spaced evenly between $0°$ and $90°$. The descriptor output $\mathbf{v}$ contains the values from the bins of the histogram, normalised such that $\|\mathbf{v}\| = 1$. Hence nb also determines the dimensionality of the final descriptor.

SIFT and SURF involve an orientation assignment step that makes the rest of the process invariant to rotation. Because THRIFT uses only a comparison of surface normal estimates at two scales, the descriptor is already invariant to full 3D rotation, and there is no need for an explicit orientation assignment step.

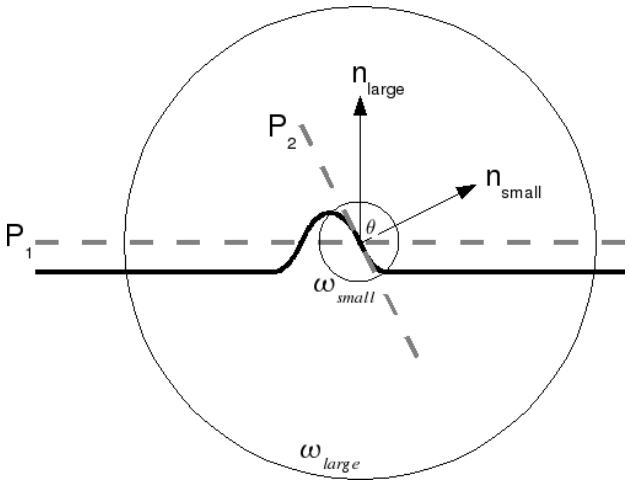

Figure 2. The two least squares planes and corresponding normals for one support point on an example surface

4. Results

We evaluated ThrIFT using data acquired with a laser range finder. Each data set contained approximately 500 million data points collected in a $360°$ sweep.

Since the primary goal of the detector is to be repeatable, our first experiment tested the repeatability of our detector in the presence of noise. We generated a density map from the range data and then added independent Gaussian noise to each density value. We ran the detector before and after the addition of noise and recorded the number of interest points that were detected in both cases as a percentage of the total number of interest points. We used density maps of size $100 \times 100 \times 100$. Our detector generated 5 octaves, each containing 4 layers. We used a threshold of 0.5. Only those interest points that were detected at exactly the same location and scale were classified as repeats.

Figure 3 shows the observed repeatability as a function of noise. The results show a high level of repeatability. At 1% noise, our detector achieves a repeatability of 91.3%. At 10% noise, which corresponds to less than 3 bits of precision, our detector repeatably detects over 50% of the interest points at exactly the same location and scale. In the context of extrapolating a partial model of a scene (which is the intended application of ThrIFT) there will be many features in the scene that could be matched. Detection of any significant subset will be enough to extrapolate the model;

hence these results show that ThrIFT is suitable for this application.

We evaluated the performance of our descriptor on the problem of detecting repeated structure in 3D scenes, since that is the key information needed to extrapolate a partial model. We used urban scenes containing buildings with repeated structure. For each scene we used the detector to find a set of interest points. We then computed the descriptor for each interest point and compared these with the descriptor computed for a hand-picked reference region. We isolated the 50 interest points with descriptors that most closely matched that of the reference region, and recorded how many of these corresponded to repetitions of the scene structure at the reference region (correct matches). Matching was performed using the Euclidean distance between descriptors. Ground truth was established manually.

We used 10 bins for the histogram. We set $\omega_{small} = 0.3\sigma$ and $\omega_{large} = 0.8\sigma$ where σ is the radius of the reference region. Hence ω_{small} and ω_{large} were constant for all interest points.

Table 1 shows the results of this experiment for three data sets. Figure 5 shows the locations of the specific regions that were matched to the reference point in each scene. The results show that of the strongest 50 matches, over 80% in each scene were correct identifications of repeated structure.

The first two test scenes, "Library-Sparse" and "Windows-Sparse", contained facades that were oblique to the range finder, resulting in sparsely sampled surfaces (see Figure 5). For these scenes ThrIFT was still able to achieve an accuracy above 80%. The last scene, "Windows-Front" contained a more densely sampled facade, which led to a corresponding increase in accuracy (94%).

In this evaluation we considered each interest point and descriptor independently. A more intelligent use of the available data would be to consider the descriptors together using some higher-level recognition system, which would lead to better detection of repeated structure. For example, we might look for many regularly-spaced matching descriptors as evidence for repeated structure, or we might look for repeated groups of descriptors.

We conducted our experiment without any such higher-level integration of the information so that our results would show the performance of ThrIFT alone. Since ThrIFT was alone able to achieve such promising results, we can expect good performance when we use ThrIFT as input to a higher-level recognition system designed specifically for detecting repeated structure.

5. Conclusion

In this paper we have presented ThrIFT, a system that extracts and describes distinctive scene feature from range data. We have justified our approach as a generalisation of two successful 2D feature extraction systems, SIFT and SURF.

We have shown the performance of our system by testing it on a number of scenes acquired using a laser range finder. Our results show that our detector exhibits high repeatability, and that our descriptor can be used for identification of repeated structure.

Future work will focus on more comprehensive evaluations of ThrIFT. We will test the repeatability of the detector in the presence of viewpoint changes, and systematically test the descriptor on a library of range data scans.

Data set	% Correct Matches
Library-Sparse	86%
Windows-Sparse	84%
Windows-Front	94%

Table 1. Out of the strongest 50 matches, the percentage that were correct

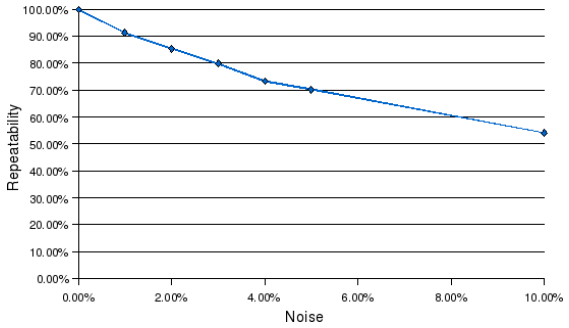

Figure 3. Repeatability of the interest point detector in the presence of noise.

References

[1] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *Ninth European Conference on Computer Vision*, May 2006.

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

[3] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):433–449, 1999.

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

[5] A. van den Hengel, A. Dick, T. Thormählen, B. Ward, and P. Torr. Videotrace: Rapid interactive scene modelling from video. In *SIGGraph Submission ID 425, to appear*, 2007.

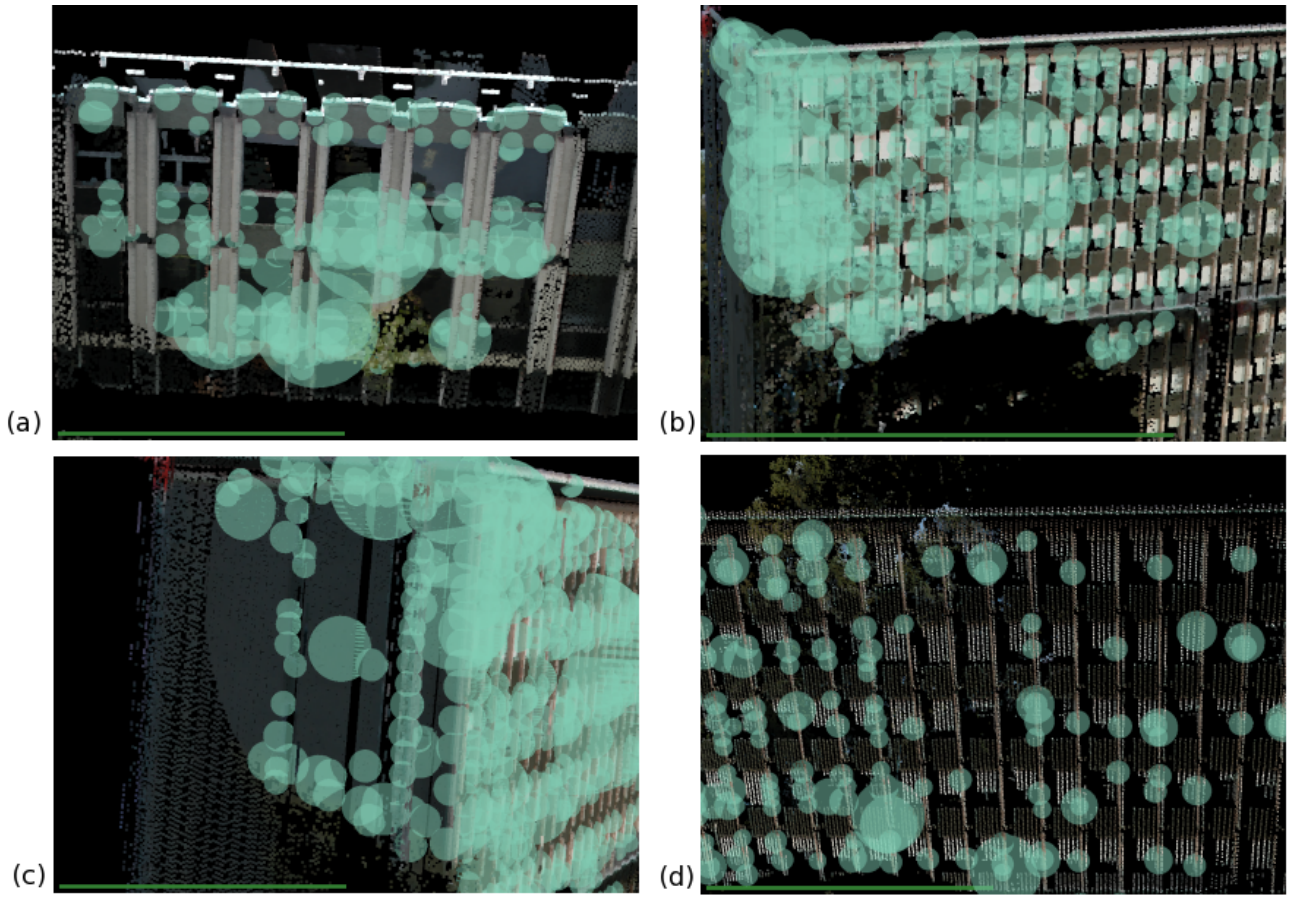

Figure 4. Results from the detector. The spheres show the location and scale of all detected interest points. Notice how interest points tend to occur in regions with high spatial gradients.

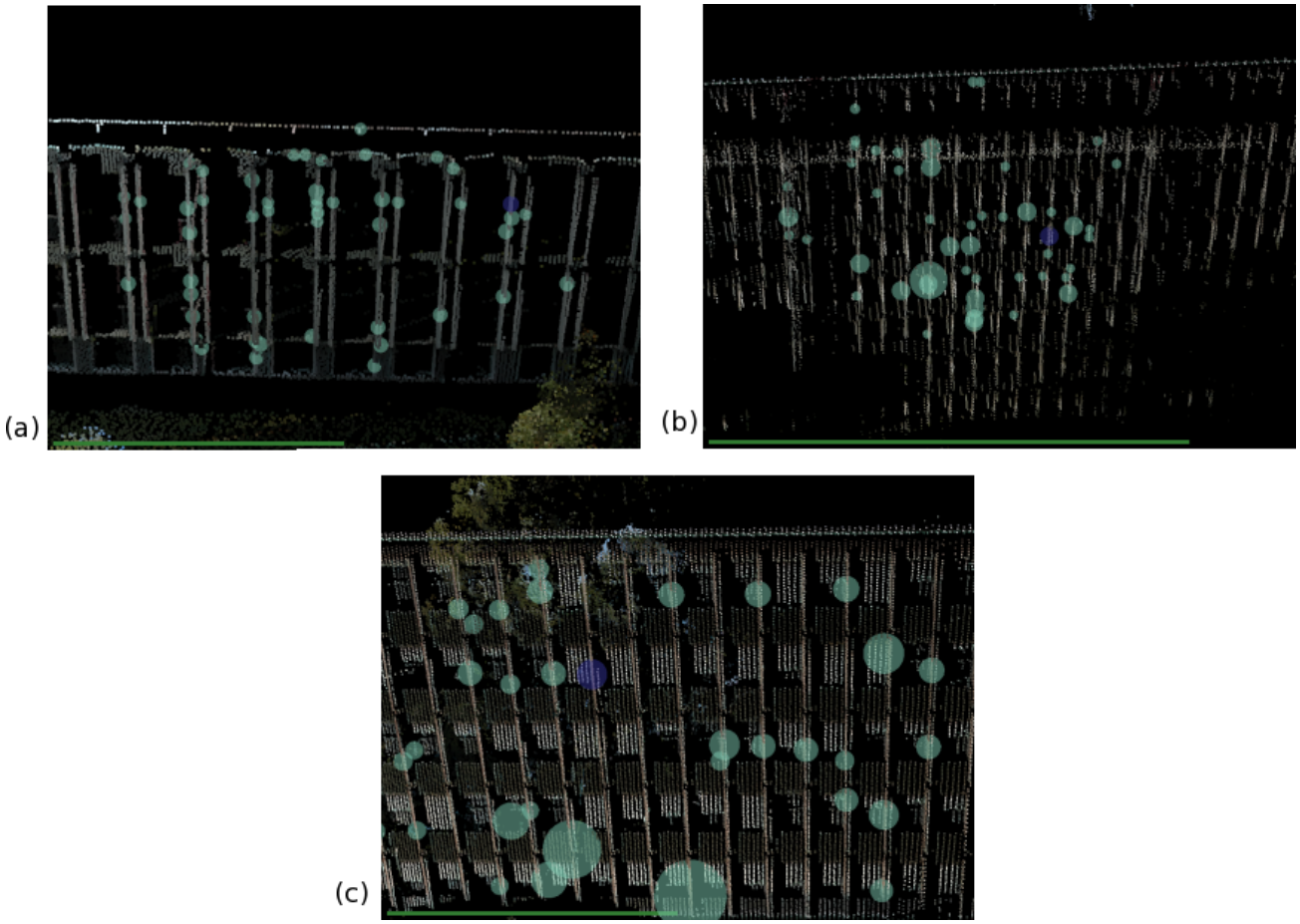

Figure 5. Results from the descriptor. The blue sphere shows the location and scale of the reference point. The green spheres show the locations of the best 50 matches. Even from an oblique viewpoint, repeated structure is identified in the majority of cases. The data set names (for cross-reference with Table 1) are (a) Library-Sparse; (b) Windows-Sparse; (c) Windows-Front.