3D Target Recognition Using 3-Dimensional SIFT or Curvature key-points and Local spin Descriptors

D. Gibbins

Sensor Signal Processing Group, Department of Electrical & Electronic Engineering,
The University of Adelaide, Australia.
Email: danny@eleceng.adelaide.edu.au

Abstract

The recognition of a target object from 3D range data has applications in aerial surveillance and robotics including the controlling of intelligent autonomous UAV's and UGV's. This paper describes a 3D target recognition scheme based on modified versions of 3D-SIFT and local curvature maxima techniques originally proposed by Flint et.al[1] and Ho et.al.[2, 3] The modified schemes use local SPIN descriptors, which are partially invariant to rotation, at key points across various scales to identify common structures between a reference 3D target and a scene imaged by a range sensor. The identified common structures are clustered locally to form transformation estimates and associated match hypotheses which are then tested to identify the most likely location of the target object in the scene. By limiting the search to those denoted by clusters of local matches, the search space of possible solutions is greatly reduced.
The results of applying this method to simulated range data are presented to demonstrate the proposed approach.

1 Introduction

The recognition of a target object from 3D range data of a reference object and imaged scene, such as that shown in Figures 1 and 2 has applications in aerial surveillance and in the controlling of intelligent autonomous UAV's and UGV's. Numerous approaches to recognition from 3D have been investigated such as appearance, exhaustive search and local descriptor based approaches[4, 5].

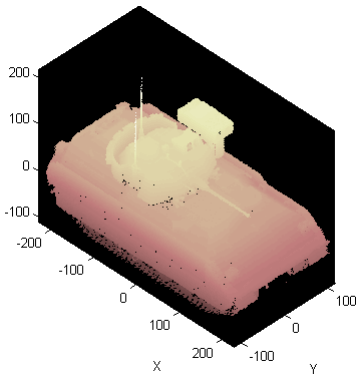

Figure 1: A 3D reconstruction of an M2A2 based on simulated data.

This paper proposes a 3D target recognition scheme based on modified versions of 3D-SIFT or local curvature maxima techniques originally proposed by Flint et.al[1] and Ho et.al.[2] to identify key-points

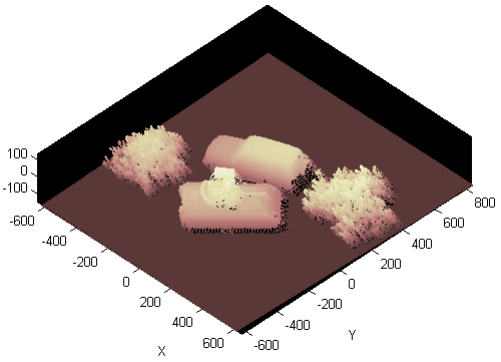

Figure 2: A 3D reconstruction of a scene containing an M2A2 and other structures as might be observed from a low-flying UAV (simulated data).

in 3D point-cloud data. This work attempts to combine these ideas with spin image descriptors originally proposed by Johnson[5] which are used to perform key-point to key-point matching. A target matching approach is proposed which uses local clusters of spin image matches to identify possible poses of the target object in the scene which can then be tested. By limiting the search to triangular clusters of local matches, instead of comparing all possible combinations, the search space of possible solutions can be greatly reduced.

The remainder of this paper is laid out as follows. Firstly the previous work of Johnson[5], Flint[1] and

Ho[2] is described. Based on this work this paper then describes a target recognition scheme which attempts to use local clusters of spin image descriptors at key-points to determine possible target matches. The results of applying this method to simulated range data are then presented to demonstrate the proposed match approach and a comparison is made of match performance using 3D-SIFT or curvature based key-points.

2 Previous Work

2.1 Spin Image Recognition

Spin images, originally proposed by Johnson[5], are a well known method of encoding local surface structure into a series of two dimensional representations and have been used widely in 3D object recognition applications[4]. For a given point p on the object the local surface is parameterised by two coefficients α and β representing the radial distance and depth relative to the point of interest and the local surface normal estimate. That is:

$$\begin{aligned} \alpha(x,y,z) &= \sqrt{x^2+y^2} \\ \beta(x,y,z) &= z \end{aligned}$$

where the coordinate system is centred on p and is oriented with respect to the local surface normal. The spin image of this local patch is a discrete approximation to the PDF of the surface in terms of these two coefficients.

An illustration is given in Figure 3.

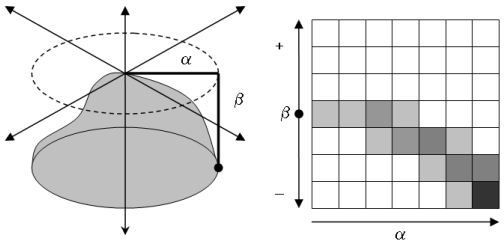

Figure 3: A diagram illustrating the local spin image coordinate system.

The resulting spin-image features are invariant to rotation (spin) of the surface about the local normal direction and encode much of the local surface structure. Moreover spin images from different models can be readily compared to one another using correlation.

The main issue with the spin image features for target recognition as originally proposed in [5] is in finding combinations of features in the model which lead to a global match solution with the test data, if one exists. Two approaches used in previous work [4] are to either apply an exhaustive search, which may be impractical, or to test sufficient random combinations of points to maximise the likelihood of a match. This latter approach, however, does not guarantee that any correctly matching combination of points will be tested.

2.2 3D-SIFT (THRIFT) Matching

3D-SIFT is a 3-dimensional variant of the SIFT region matching approach proposed by Lowe[6] for finding correspondences in 2D imagery. In 2D-SIFT [6] a smoothing filter is applied to an image at varying scales. Local maxima in the differences between these smoothed images is used to identify key points which can be identified in a similar scene independently of the scale, position and orientation. In [6] each point is represented by an elaborate description of the directional distribution of the local image gradients around the key-point which is used for feature point matching and ultimately image registration or object recognition.

In Flint[1], a 3D variant of this 2D approach (called THRIFT) was proposed for the analysis of 3D point-cloud data generated by LADAR or similar range sensing systems. In [1] the 3D data is converted to a 3D array of rectangular subregions (ie. voxels) represented by a normalised density function of the form:

$$D(i,j,k) = \frac{n\left(B(i,j,k)\right)}{\text{argmax}_{(i,j,k)\in I}\left\{n\left(B(i,j,k)\right)\right\}} \qquad (1)$$

where $n(B(i,j,k))$ is the number of range data points within the voxel $B(i,j,k)$.

Regions of interest in this density function are then identified by applying the determinant of the Hessian $H(\hat{\mathbf{x}},\sigma)$ defined as

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

where

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

and $g(\sigma)$ is a 3D Gaussian of variance σ. By computing the Hessian at multiple scales and searching both in position and scale for local maxima, a set of key-points X can then be identified, where

$$\text{interest}(X) = \text{arglocalmax}_{\hat{\mathbf{x}},\sigma} \left|\det\left(H(\hat{\mathbf{x}},\sigma)\right)\right| \qquad (2)$$

Flint et.al [1] went on to define a local invariant feature based on angular variations in the local surface normal directions as a possible method for point matching between two scenes. Their work however stopped well short of developing the approach for 3D target recognition.

An examination of THRIFT has highlighted two drawbacks with their approach. Firstly the density function is sensitive to regions of overlapping data, and secondly their proposed normal variance description of the surface, whilst invariant to rotation, captures too little of the surface structure leading to mismatching with other types of surface shape.

2.3 Local Curvature Matching

Local surface shape can be described in terms of its principal curvatures k_1 and k_2. For a 3D point cloud the curvature at each point can be calculated in several ways such as fitting a 2nd order polynomial approximation using points within a local neighbourhood [7]. Varying the size of the local neighbourhood around each point provides a mechanism for estimating the local curvature at different scales. From k_1, k_2 several different descriptors of the local curvature are often used. These include the principal curvature and shape index defined as

$$C(\hat{\mathbf{x}}, n) = k_1(\hat{\mathbf{x}}, n)k_2(\hat{\mathbf{x}}, n)$$

and

$$s(\hat{\mathbf{x}}, n) = \frac{\pi}{2} \arctan\left(\frac{k_1(\hat{\mathbf{x}}, n) + k_2(\hat{\mathbf{x}}, n)}{k_1(\hat{\mathbf{x}}, n) - k_2(\hat{\mathbf{x}}, n)}\right)$$

respectively. Here $(\hat{\mathbf{x}}, n)$ represents the position and scale at which the curvature components were computed.

In [2] local maxima of the principal curvature across position and scale were used as points of interest, that is:

$$\text{interest}(X) = \text{arglocalmax}_{\hat{\mathbf{x}},n} C(\hat{\mathbf{x}}, n) \qquad (3)$$

As with [1], the resulting principal curvature or shape index values are independent of the position and orientation of the surface patch. By selecting points where the curvature is locally maximal in space and scale, it is possible to identify the same points on the surface of an object independently of the object's scale, position or orientation.

In later work by Ho et.al [3], these local curvature estimates were used to match local patches of similar curvature with consistent spatial separation between two object models leading to a curvature based target recognition scheme. One potential limitation of the approach in [3] is the heavy reliance on the spatial separation to identify suitable match combinations amongst the large number patches which may possess very similar curvature characteristics. Another drawback relates to the computational requirement of estimating curvature of a surface at each point.

3 The Proposed Approach

3.1 Feature Key-Point Identification and Descriptors

In the key-point estimation approach proposed here the 3D-SIFT key-point detection approach of [1] is applied to a modified version of the point density function $D'(i, j, k)$ where

$$D'(i, j, k) = \begin{cases} 1 & \text{if } n(B(i, j, k)) > t \\ n(B(i, j, k))/t & \text{otherwise} \end{cases}$$

where t may be as little as 1. This modified version of the density function is less sensitive to variations in surface sampling density caused by either the sensor observation conditions or overlapping segments of the range data from repeated scanning.

An illustration of this stage of the processing for the Hessian key-point features applied to the M2A2 model is shown in Figure 4. Here the Hessian was computed at 5 different scales between 1.0 and 2.5 using a voxel size of 4cm. In this example a model of some 100k voxels has been reduced to a set of around 400 key-points.

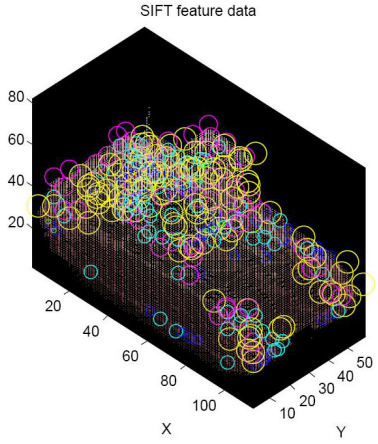

Figure 4: Example of sift detections for the M2A2 example shown earlier. The size of the circles is an indication of relative scale of the identified key-points.

For the equivalent curvature key-point estimation of [3] this representation of the point density is also used to define the points at which curvature is estimated.

Given the key-point locations determined by the above processes, local spin images are then computed at each location at the same relative scale as the key-point. For the purposes of the experiments presented herein, a local 11×11 spin image descriptor was used to represent each key-point location.

3.2 Matching

The matching of the 3D reference model to the 3D scene data requires us to hypothesise and test the likely pose of the object based on local correspondences between the estimated model and scene features. In the proposed approach, the match step is performed by identifying combinations of local matches between the spin features in the scene and model data which are geometrically consistent. These are then used to hypothesise a pose for the target in the scene which can be tested using a likelihood match score.

In this work, motivated by [3], a Delaunay triangulation of key-points is used and combinations of features forming the facets of the triangulation are used for matching. This use of local groupings of features rather than exhaustive or random searching significantly reduces the search space whilst preserving some robustness to partial occlusion of the target. However, unlike [3], the starting combinations of key-points used here is further restricted to key-points in the model data with high correlation to features in the scene data, and vice versa for the scene data.

Next, a set of local feature combinations for detailed testing is then identified by examining each triangulation in the model data and identifying a structurally similar triangle of features in the scene data where the spin feature locations are both spatially consistent (ie. the ratio of their distances is similar) and their spin representations are close matches. In this way the initial set of combinations is further pruned to those most likely to form coherent matches with the scene. This change to [3] also avoids the situation of high correlation features being grouped with nearby features with low correlation to structures in the scene. The resulting identified pairings of target and model features are then used to estimate a target pose which is tested using a maximum likelihood estimator.

Based on this, the match approach proposed here can be summarised as follows:

Algorithm:

1. *Find high correlation matches between the model and scene spin feature measures*
2. *Form a Delaunay triangulation of all points in the model with good matches to the scene data*
3. *Repeat this process for the scene points with good matches with the model data*
4. *Identify pairs of triangles in model data and scene data that have consistent size and shape*
5. *For each identified pair compute the transformation between them and use this to project the model data into the scene and from this determine a match score*
6. *Identify the match as the combination with the highest score exceeding a given match tolerance*

In the above algorithm the recognition score is based on the alignment error between the two nearest key-points and the correlation score of their spin features. For each key-point a likelihood function of the form

$$m(d, c) = e^{-\frac{d^2}{\sigma_d^2}} e^{-\frac{c^2}{\sigma_c^2}} \tag{4}$$

can be used where d and c are the distance and associated match scores between the nearest model and feature points after projection. The parameters σ_d and σ_c are user defined tolerances on the distance and match errors respectively.

Once a target match is identified, a finer registration of the target and model data can be achieved using the iterative closest point (ICP) algorithm[8].

4 Preliminary Experimental Results

To assess the proposed recognition approach, two simple models from the Stuttgart Range Image Database[9] were employed to assess matching and registration in the absence of other object clutter. This database contains a series of range images of objects viewed from multiple different directions. For this experiment a representation of the object seen from one viewing direction was compared with another view of the same object.

Next, a series of more realistic simulated scenes containing the M2A2 model shown earlier and other clutter targets was constructed under different viewing conditions (refer to the example in Figure 2). In this case a completed 360 degree reconstruction of the M2A2 was used as the source data.

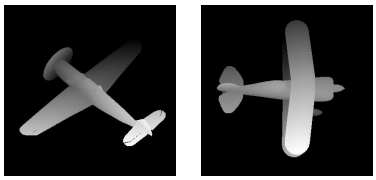

Figure 5: Example range images of two test objects from the Stuttgart Range Image Database.

For the purposes of evaluation the M2A2 scene data was converted to a $4 \times 4 \times 4$cm voxel representation and the SIFT and curvature features computed over 5 different scales representing regions of radius 7 to 29 voxels. A similar sampling resolution relative to the size of the test objects was used for the Stuttgart range imagery.

Both Hessian and curvature based key-points were estimated from the model and scene data. The local spin images were represented as 11×11 images at the same position and scale as the detected key-point. It should be noted that estimation of the curvature key-points required significantly greater calculation time than the SIFT key-points.

4.1 Simple Aeroplane Models

Tables 1 and 2 present the results of applying the matching approach to a set of 15 range images of each model from the Stuttgart Range Image Database[9]. Here individual range views of the object were compared to each other to test the ability of both the key-point estimation and match stages for the Hessian and a curvature maxima key-point features. The model and reference data were considered to be a good match if the angular errors were less than 10 degrees and the relative positional error was less than 5% of the target's size.

Overall, the Hessian based key-points tended to be more reliable for matching and achieved 100% correct matching for both models. The poorer overall alignment and observed failure in the curvature based match for the fighter model appears to be related to a lack of detected key-points near the wing tips of the aircraft models.

Key Point	Match	Mean Abs Error			
		d	h^o	p^o	r^o
Hessian	100%	3.4	1.3	0.9	1.3
Curvature	93%	8.6	3.6	4.8	2.8

Table 1: Results for the "Fighter" model data where d is the distance error, and h^o, p^o and r^o are the heading pitch and roll errors respectively.

Key Point	Match	Mean Abs Error			
		d	h^o	p^o	r^o
Hessian	100%	4.3	1.7	1.2	0.9
Curvature	100%	5.1	1.6	1.1	1.1

Table 2: Results for the "Pitt (bi-plane)" model data.

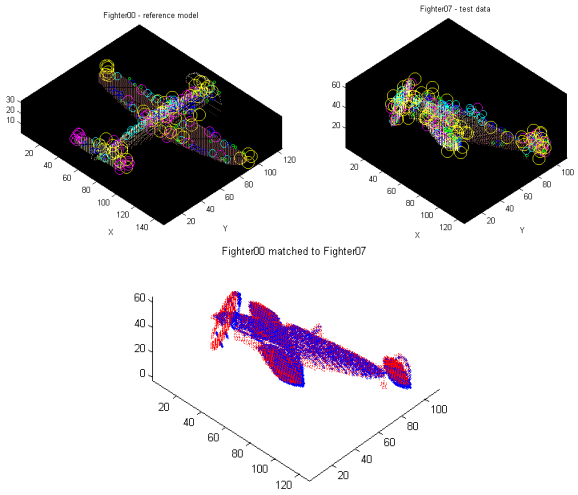

Figure 6: Example SIFT feature and match result for the Fighter model data SIFT key-points. In the lower plot, the model and test scene are shown in blue and red respectively.

4.2 M2A2 Recognition Results

Table 3 shows the results of comparing the Hessian based key-point matching to key-points based on local curvature information for a set of 16 simulated range images of a scene containing the M2A2 model, a 4WD vehicle and simulated vegetation taken at different viewing angles. As can be seen both the Hessian and curvature features resulted in successful matches with angular errors of typically less than 2 degrees. An example match result is shown in Figure 7. Comparisons with curvature variance and shape-index variance suggested for key-point identification in [2] are also shown.

Whilst the proposed recognition scheme performed perfectly in these tests, again the Hessian key-points seemed to produce slightly smaller registration errors. The curvature variance and shape index key-points did not produce as good results.

5 Conclusions

This paper proposes a 3D target recognition scheme based on a combination of ideas from [1, 2, 5] and local clusters of matching features. Preliminary results of applying this work to simulated range data

Key Point	Match	Mean Abs Error			
		d	h^o	p^o	r^o
Hessian	100%	2.9	0.8	0.5	1.1
Curvature	100%	4.8	1.9	1.0	1.4
C variance	88%	21.2	4.1	9.4	17.1
SI variance	100%	6.9	3.5	1.5	3.2

Table 3: Results for the "m2a2" scene data.

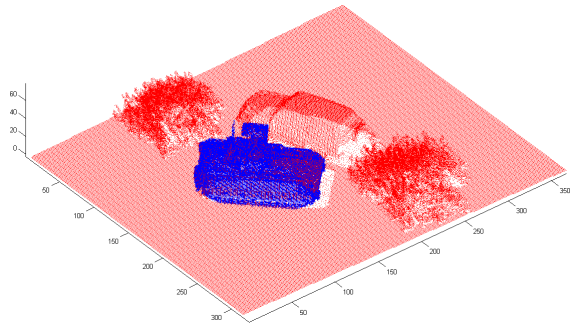

Figure 7: Example match result for the M2A2 model and scene using SIFT key-points. The model and test scene are shown in blue and red respectively.

using Hessian or curvature based key-points were presented in Section 4. These results demonstrate that this approach is capable of successfully performing target recognition and registration for 3D targets for the test scenes considered here. Based on the comparative results of the Hessian and Curvature key-points, it would appear that the Hessian key-points are more likely to result in a good target match. Given the relatively lower computational requirements of the Hessian versus local curvature computations, these would appear to be better suited for real-world application.

Future work on this proposed approach will focus on a more thorough assessment of match performance on real data and scenes involving partial occlusion of the target object.

6 Acknowledgements

The author would like to acknowledge the sponsorship of Dr Leszek Swierkowski and Dr Anthony Finn (DSTO Edinburgh, Australia) which originally motivated this research work, and the assistance of Mr Michael Driscoll (DSTO Edinburgh, Australia) in the generation of the 3D simulated vehicle data.

References

[1] A. Flint, A. Dick, and A. van den Hengel, "Thrift: Local 3d structure recognition," in *Proc. Digital Image Computing Techniques and Applications*, pp. 182–188, December 2007.

[2] H. Tho and D. Gibbins, "Multi-scale feature extraction from 3d models using local surface curvature," in *Proc. Digitial Image Computing: Techniques and Applications (DICTA'08)*, (Canberra, Australia), December 2008.

[3] H. Tho and D. Gibbins, "Multi-scale feature extraction for 3d surface registration using local shape variation," in *Proc. Image and Vision Computing New-Zealand (ICVNZ'08)*, (Christchurch), November 2008.

[4] R. Campbell and P. Flynn, "A survey of free-form object representation and recognition techniques," *Computer Vision and Image Understanding*, vol. 81, pp. 166–210, Feb 2001.

[5] A. Johnson and M. Hebert, "Using spin images for efficient object recongition in cluttered 3d scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, pp. 674–686, May 1999.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Internaltional Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] J. Goldfeather and V. Interrante, "A novel cubic-order algorithm for approcimating principal direction vectors," *ACM Transations Graph*, vol. 23, pp. 45–63, Jan 2004.

[8] S. Rusinkiewwicz and M. Levoy, "Efficient variants of the icp algorithm," in *Proceedings of 3DIM*, pp. 145–152, 2001.

[9] "Stuttgart range image database," http://range.informatik.uni-stuttgart.de/htdocs/html/: Stuttgart University.