Point-cloud Registration Using 3D Shape Contexts

Mathew Price
Cogency cc
Cape Town
Email: mathew@cogency.co.za

Jeremy Green
CSIR
Centre for Mining Innovation
Johannesburg
Email: jgreen@csir.co.za

John Dickens
CSIR
Centre for Mining Innovation
Johannesburg
Email: jdickens@csir.co.za

***Abstract*—The problem of aligning scans from a range sensor is central to 3D mapping for robots. In previous work we demonstrated a light-weight descriptor-based registration method that is suitable for creating maps from range images produced by devices such as the XBOX Kinect. For computational reasons, simple descriptors were used based only on the distribution of distances between points. In this paper, we present an alternative approach using 3D Shape Contexts that also retains angular information thereby producing descriptors that are more unique. Although this increases the computational load, intrinsic properties of the descriptor facilitate keypoint selection, leading to a more robust registration framework. This also provides greater flexibility when applying the method to sparse point clouds such as those produced by laser range scanners. Results are shown for registering new data acquired from an underground mine environment.**

I. Introduction

A customised 3D thermal mapping sensor [1] is currently being developed as part of the CSIR's mine safety platform [2]. This involves generating a robo-centric map of the mine ceiling and texturing the model with thermal imagery acquired from an infrared camera. The resulting map can then be used in conjunction with other sensors to assess risk. Since the thermal mapping sensor is self-contained and independent of the robot's navigation sensors, the problem of generating a registered map from sequential scans must be tackled. Having a non-integrated sensor is beneficial as it also allows the unit to be used as a hand-held scanner for data gathering and research.

In previous work [3] we presented a scan registration algorithm based on light-weight descriptors dubbed *Distance Signatures* (dSig). This was selected over competing methods such as 3D Shape Contexts [4] and Spin Images [5] due to processing requirements, and matching inefficiencies. However, since the latter methods incorporate angular information their descriptors can be more uniquely identified, which leads to greater robustness during registration.

Recent development focusing on improving mapping for longer sequences has led to an alternative scan registration algorithm based on 3D Shape Contexts (SC3). In this paper, we discuss how some of the previously identified issues have been addressed and show how the new descriptors offer greater flexibility.

II. Sensors

The mapping sensor is equipped with off-the-shelf hardware, and in particular 3D points are acquired from an XBOX Kinect. This makes it cost-effective to construct multiple setups and reduces the mechanical complexity required to manipulate line-based laser scanners. The Kinect comprises a colour camera and an infrared camera-projector pair, and produces calibrated range data.

More recently, we have also used the Asus Xtion that is based on the same technology, but is more compact (both are derived from PrimeSense's *PrimeSensor*). Since these devices (shown in Figure 1) produce 2D range images, standard camera calibration techniques can be applied to determine relative poses with respect to other sensors, which is very convenient. For instance we register a 3D sensor with a thermal imaging camera to provide temperature-textured 3D maps.

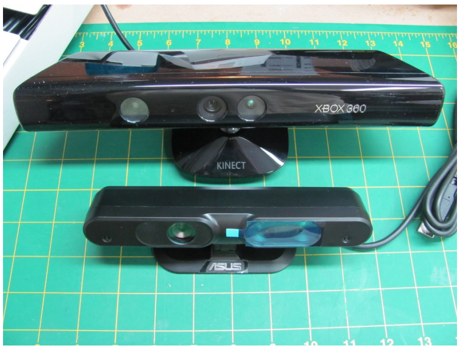

Fig. 1. *Top: Microsoft XBOX Kinect; Bottom: Asus Xtion Pro Live.*

Although the Xtion and Kinect have much lower operating ranges than conventional laser scanners, our application only calls for generation of localised ceiling maps which does not require ranges above several metres. (The Kinect can measure up to 8m in unlit environments.) Another factor is that the subtle thermal readings that are of interest are more reliable when measured at close proximity. Extended maps are generated by stitching multiple sets of locally registered scans. We have gathered data from several underground mine stopes and found that the sensors are well suited to the unlit dusty environment.

III. Registration with Descriptors

Computing a registered 3D surface with descriptors is closely related to the method used to create panoramas from 2D photos, often called image stitching [6]. Interesting points (keypoints) are detected in each image; local descriptors are computed around each keypoint, and subsequently used to find corresponding matches in other images. Robust parameter estimation is then used to estimate the relative pose between image pairs, and by choosing a common reference frame a stitched image is created. This methodology has shown to be very effective for 2D data since: it provides once-off detection for individual images (which is efficient); robust estimators can be used (e.g. RANSAC) for removing outliers; and it can be applied to sparsely captured and partially matching data. (i.e. no assumptions are made about the motion of the sensor). The aforementioned methods are also invariant to image intensity, rotation and scale which are desirable properties for descriptors.

A. 3D Descriptors

Several 3D descriptor methods have been proposed for stitching 3D data in a similar way [7], [8], [4], [9]. A primary requirement of an interest-point detector is that similar keypoints are chosen for each image. Furthermore, the set of keypoints should be efficiently selected so as to produce a compact yet ample supply of unique descriptors allowing efficient matching.

In 2D, methods such as SIFT [10] and SURF [11] achieve this by generating a multi-resolution image pyramid and selecting keypoints that are consistent over a range of scales. So far, this has been difficult to apply directly for 3D range images because pixels represent distances measured to the center of the camera instead of visual properties of objects. (i.e. descriptors change with camera motion). In [12] local curvature estimates are used to construct a normalised image pyramid. Unfortunately, 3D range images from the Kinect are prone to variations that cause curvature to be an unreliable measure. In addition, constant variation of the projector's pattern causes spurious missing data patches that alter the local image. Therefore, we focus on the raw point data for features rather than range-image extensions of 2D methods.

B. Keypoint Selection

Currently, there is no widely adopted method for selecting repeatable keypoints for 3D descriptors. Therefore, the fall-back of random or uniform sampling is often used, which relies on the uniqueness of the descriptors. In object recognition applications, this is acceptable since the exact location of matching keypoints in candidate pairs can be relaxed. For instance, descriptors for two candidate keypoints (one from the reference scan and one from the query scan) that are close to one another will be similar resulting in a match. However, with registration the positions are also used to compute the relative alignment, so locations are critical for high accuracy. In [3] we exploited the sequential nature of the data where only small motions were detected between successive frames. Using a uniform keypoint selection process circumvented the problem at the cost of extra processing, which was addressed by using light-weight descriptors. However, no measure of uniqueness was available for filtering similar descriptors, and this resulted in confusion in scenes with insufficient 3D variation (e.g. walls, pipes, etc.). In fact, because the Kinect has relatively low 3D resolution for finer surfaces, even mine tunnels appear to be problematic when viewed head-on. For this reason, we explored the more computationally intensive option of using 3D Shape Contexts that more accurately describe the local neighbourhood. Although we still retain the uniform sampling technique, the addition of orientation sensitivity to the descriptor reduces mismatches as a result of the nearby-keypoint issue discussed.

IV. 3D Shape Contexts

Shape Contexts [13] were originally proposed for describing 2D point sets for recognising handwritten characters and embedding such objects for image-based retrieval. Essentially, they are 2D histograms (shown in Figure 2) formed by binning points surrounding a keypoint according to angle and logarithmic radius. The use of logarithmic radius provides good contextual information while reducing sensitivity to distant objects.

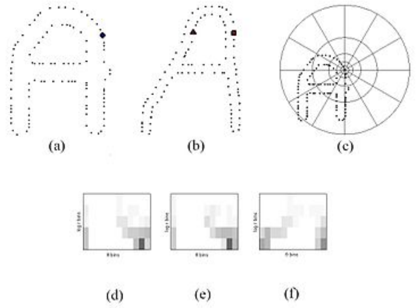

Fig. 2. *(a) Shape 1 showing keypoint u. (b) Shape 2 showing keypoint u and v. (c) Shape Context grid shown for Shape 1 u. (d) Shape Context for Shape 1 u . (e) Shape Context for Shape 2 u. (f) Shape Context for Shape 2 v. (e) and (f) are dissimilar due to their different positions.*

For 3D points, the method can be extended by binning points according to azimuth, elevation and logarithmic radius within a sphere of influence. In order to match descriptors shape contexts must be computed using the same reference frame. For 2D polygons this can be achieved using the edge normal and choosing a tangent consistent with the vertex order (clockwise or counter clockwise). However, in 3D defining a reference frame using the local surface normal still leaves an unknown rotation about the azimuth direction which cannot be uniquely determined. The conventional approach involves either using a 1D search over the unknown angle during matching (and selecting the best match), or precomputing multiple descriptors for each keypoint at different azimuth intervals. Both of these options are not efficient and reduce descriptive power; in fact this was what motivated us to propose Distance Signatures in the first place.

A. Unique Shape Contexts

Recently, Tombari et al. [9] showed that a simple yet intuitive solution proposed by Bro et al. [14] held the key to solving the problem of selecting reference frames for 3D descriptors. This lead to their definition of the Unique Shape Context [15]. The method builds on the idea of using least squares to estimate the local surface normal [16], and uses Principal Component Analysis (PCA) within a fixed radius of a keypoint to specify an orthogonal reference frame. This presents two problems: (1) sign ambiguity of the three axes and (2) angular ambiguity of the tangent plane about the normal in cases where the eigenvectors are not unique. We do not perceive the latter to be an issue since the extent to which the eigenvectors are not unique is the extent to which the points are distributed symmetrically, in which case the descriptor would not be unique. For instance, keypoints on a flat wall are a poor choice for registration since they cannot be described uniquely. The USC proposal consists of two steps: First, contributions $\mathbf{p}_i$ to the covariance matrix M are weighted according to distance in order to improve repeatability in the presence of background clutter and noise:

$$k = \sum_{i:d_i \leq R} (R - d_i) \tag{1}$$

$$\mathrm{M} = \frac{1}{k} \sum_{i:d_i \leq R} (R - d_i)(\mathbf{p}_i - \mathbf{p}_k)(\mathbf{p}_i - \mathbf{p}_k)^{\mathrm{T}} \tag{2}$$

(Keypoint $\mathbf{p}_k$ is used in place of the centroid, d_i represents the distance between $\mathbf{p}_k$ and $\mathbf{p}_i$, and R is a maximum fixed radius of interest.) Second, the sign ambiguities are resolved by selecting directions that agree with the majority of the data [14]. Let the principal axes be denoted $\mathbf{x}^+, \mathbf{y}^+, \mathbf{z}^+$ or $\mathbf{x}^-, \mathbf{y}^-, \mathbf{z}^-$ depending on the sign. Their signs are disambiguated as follows:

$$S_x^+ = \{i : d_i \leq R \wedge (\mathbf{p}_i - \mathbf{p}_k) \cdot \mathbf{x}^+ \geq 0\} \tag{3}$$

$$S_x^- = \{i : d_i \leq R \wedge (\mathbf{p}_i - \mathbf{p}_k) \cdot \mathbf{x}^- \geq 0\} \tag{4}$$

$$\mathbf{x} = \begin{cases} \mathbf{x}^+, |S_x^+| \geq |S_x^-| \\ \mathbf{x}^-, \text{otherwise.} \end{cases} \tag{5}$$

$$\mathbf{z} = \begin{cases} \mathbf{z}^+, |S_z^+| \geq |S_z^-| \\ \mathbf{z}^-, \text{otherwise.} \end{cases} \tag{6}$$

$$\mathbf{y} = \mathbf{z} \ \mathrm{x} \ \mathbf{x}. \tag{7}$$

For each keypoint, we can therefore obtain a repeatable reference frame by computing the eigenvectors of M and applying the sign disambiguation method. Sorting the eigenvalues in descending magnitude specifies the $\mathbf{x}, \mathbf{y}, \mathbf{z}$ axes from their respective eigenvectors where the $\mathbf{z}$ axis corresponds to the surface normal. (This is the direction of least variance.)

We now have the means to compute 3D Shape Context descriptors relative to a repeatable reference frame. Given two matching shape contexts (measured by the similarity of their descriptors) from different scans, the relative pose between the scans can be determined directly from the relative pose between the reference frames used to compute each shape context. Let T_n be a transform that translates the origin to the position of a keypoint $\mathbf{p}_{k_n}$, and R_n the 3D rotation that aligns the world to the local reference frame. Then relative pose P is:

$$\mathrm{P} = \mathrm{T}_1 \mathrm{R}_1^{-1} \mathrm{R}_2 \mathrm{T}_2^{-1} \tag{8}$$

This is a significant advantage since only one matching pair of shape contexts is required in order to completely specify the alignment between two scans. (Previously, we required three pairs of matching points since the Distance Signature does not require a canonical reference frame.) Naturally, incorrect alignment can occur as a result of incorrect correspondences produced by descriptor matching. Therefore, we use RANSAC to robustly estimate the relative pose.

V. Implementation

Registering a sequence follows a simple procedure:

1) Compute 3D Shape Contexts for current frame (or scan if using unstructured point-clouds)
2) Find matches with previous set of features (or initialise the set if this is the first scan)
3) Estimate alignment using RANSAC on candidates
4) Refine alignment with ICP (Iterative Closest Point).

Keypoints are selected by random sampling, but those with low spatial variance (determined by the eigenvalue corresponding to the z-axis) are removed. The most time-consuming task is determining which points lie within the radius of interest of each keypoint and computing the unique reference frame. We achieve this through approximate nearest neighbour searching using KD-Trees [17]. Since our ICP implementation already uses KD-Trees for similar queries, construction of the KD-Tree for each scan does not contribute additional overhead, and it can be reused. Once the neighbourhood has been defined, the canonical reference frame is constructed, and the shape context is generated.

Matching is based on pairwise comparison of descriptors using Euclidean distance. Once again, we leverage the fast query facility of the approximate nearest neighbour search and construct a KD-Tree for each set of shape contexts. This does not contribute much additional processing, because the number of shape contexts is much smaller than the size of the individual point-clouds for which KD-Trees are already computed. Small misalignments, that are produced by matching keypoints that are close but not in identical relative positions, are resolved with ICP [18] refinement.

For efficiency, we operate on downsampled versions of the range images as before [3]. However, unlike our previous implementation we do not automatically generate keyframes here. Instead we simply choose an adequate frame step. This was done to allow greater user flexibility for processing new data sets that may vary in quality, but does not preclude its reintegration in future work.

A. Normalisation

Since each bin is a sector of a sphere divided into logarithmic radial sections, they vary in size, as illustrated in Figure 3. Therefore, we follow the approach of [4] and normalise each

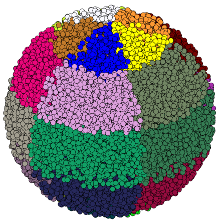

Fig. 3. *Illustration of points binned using a 3D Shape Context. Only the outer shell can be seen.*

bin by the cube root of its volume V:

$$V = \frac{2\pi}{3N_a}\left[\cos\left(\frac{w\pi}{N_w}\right) - \cos\left(\frac{\pi(w+1)}{N_w}\right)\right] \qquad (9)$$
$$\left[\left(\frac{rR}{N_r}\right)^3 - \left(\frac{R(r+1)}{N_r}\right)^3\right],$$

where w, r and R are the zero-based indices for elevation, and radius and radius-of-interest respectively; and similarly N_a, N_w and N_r are the user-specified number of divisions for azimuth, elevation and logarithmic radius. According to [4], using the cube root of the volume retains sufficient discriminative power while adding robustness against quantisation noise.

VI. Results

Through experiments with real data captured in underground gold and platinum mines, we have generated several 3D maps of areas-of-interest. Since we are interested in generating camera-centred models (like Google street-view) most of our scans do not comprise much camera translation. However, the proposed method is applicable for any type of camera motion.

A. Distance Signatures vs 3D Shape Contexts

In Figure 4 we show a comparison between models generated using our previous method based on Distance Signatures (dSig) and the new proposed method based on 3D Shape Contexts (SC3). Figure 5 shows the same comparison, but with mesh visualisations. Colours in the figures correspond to distance to the first camera with blue representing nearby points and red representing distant points.

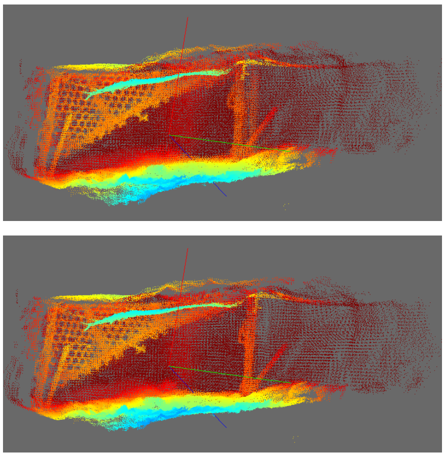

Fig. 4. *Registered 3D map inside an active mine stope. Top: Using Distance Signatures. Bottom: Using 3D Shape Contexts. Points are coloured by range to the first camera where blue is 0m and red is 2m.*

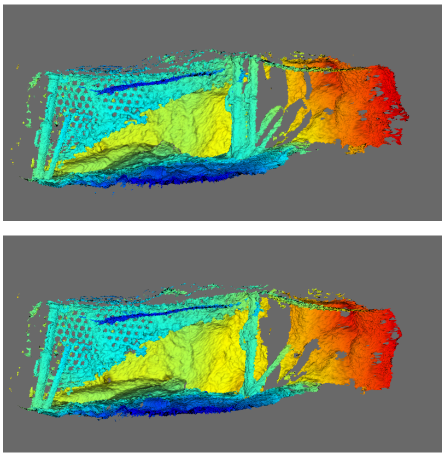

Fig. 5. *Mesh visualisation of 3D map inside an active mine stope. Top: Using Distance Signatures. Bottom: Using 3D Shape Contexts. Points are coloured by range to the first camera where blue is 0.5m and red is 3m.*

In terms of alignment accuracy both methods have similar performance, but dSig offers a 2x computational speedup. (In this example, 28s compared with 74s for SC3 on an i5 processor). However, SC3 is much more robust in general. When local variation falls below a threshold, SC3 is able to prioritise unique features and ignore non-informative keypoints; in the worst case failure is reported, whereas dSig gives no preference to uniqueness resulting in accumulated errors. An example of this can be seen in the figures where repetitive scanning of the central pillar produces poor definition and causes subsequent scans to be misaligned. However, SC3 is able to produce correct alignment.

B. Modelling a Mine Ceiling

Figure 6 shows a model of a mine ceiling generated using SC3 from a recent data set. The camera is angled upwards towards the ceiling and is rotated about its y-axis. The images show the model viewed from below.

This data was a motivating factor in developing SC3 registration since the dSig method was unable to consistently register the scans due to descriptor confusion. All the surfaces are relatively flat and descriptors taken along the pipe as the camera rotates are difficult to identify uniquely. Because SC3 incorporates angular information and can operate with very few features (recall that only one inlier match is required to propose a candidate alignment), it produces superior results.

C. Discussion

It should be noted that we have previously experimented with adding a similar variance thresholding technique to dSig, but results were still suboptimal compared with SC3. The fact that dSig offers significant computational advantage means that we have not discounted the possibility of further

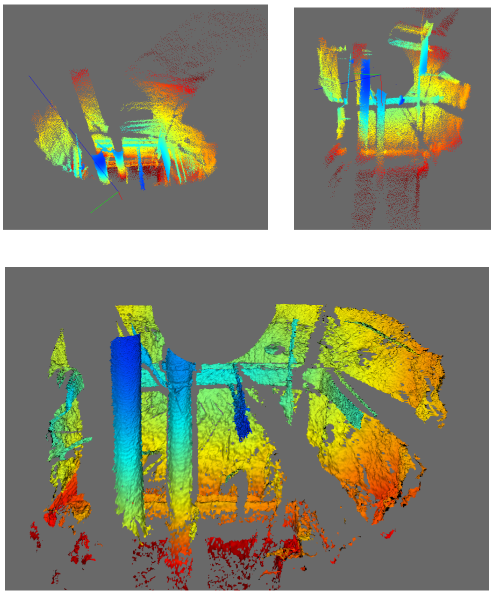

Fig. 6. *180 degree scan of a mine ceiling registered using 3D Shape Contexts. The top row shows the aligned point-cloud from different views, while the bottom row shows a mesh visualisation. Colour represents distance to the origin in the range 0.5m to 4m.*

development, though optimisation of the SC3 implementation is also a possibility.

Noisy point-clouds are a problem in extreme cases due to the reliance surface normals. (A fact we highlighted in our previous paper.) Fortunately, the PCA-based reference frame is fairly robust, especially using the distance-weighting approach. As a secondary measure, we use very few SC3 divisions (3 to 5 bins per parameter); a high number of divisions increases sensitivity to noise and processing time.

VII. Conclusions

We have presented an alternative registration method for aligning point-cloud data obtained from 3D sensors, such as the XBOX Kinect, Asus Xtion, and laser-based scanning devices. In previous work, we proposed a novel simplified descriptor that takes advantage of slow-moving sequential range images. Here we describe a more generic method based on 3D Shape Contexts, which incorporates a new technique for obtaining repeatable reference frames that improves efficiency and robustness.

Shape Contexts have previously been proposed for generating efficient shape features for 2D object recognition. More recently, they have been extended to 3D data, but suffered from ambiguities that increase processing and reduce descriptive power. By incorporating an intuitive idea that has been leveraged in data analysis problems, 3D object recognition researchers have shown how these ambiguities can be resolved in practical situations. While the concept of using 3D Shape Contexts for point-cloud registration is not novel, combining the aforementioned results to produce a generic 3D registration framework is a relatively new idea.

Comparing models generated using both descriptor systems shows that while Distance Signatures appear to be much faster to compute, 3D Shape Contexts are more reliable for automatic registration. Future work will seek to optimise the current implementation towards building a fast mapping framework.

Acknowledgements

We are grateful to the CSIR SRP for their continued support of this work, and thank the mining community for allowing us access to the various underground facilities.

References

[1] J. Dickens and M. Price, “The design of an automated 3D-thermal mine scanning tool,” in *Robmech 2012*, 2012.

[2] J. J. Green, P. Bosscha, L. Candy, K. Hlophe, S. Coetzee, and S. Brink, “Can a robot improve mine safety?” in *25th International Conference on CAD/CAM, Robotics and Factories of the Future (CARsFOF)*, 2010. [Online]. Available: http://hdl.handle.net/10204/5022

[3] M. Price, J. Dickens, and J. Green, “Creating Three-Dimensional Thermal Maps,” in *Robmech 2011*.

[4] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik, “Recognizing Objects in Range Data Using Regional Point Descriptors,” *Current*, vol. 1, pp. 224–237, 2004. [Online]. Available: http://www.springerlink.com/index/D4UKQ18FPBFE9LA6.pdf

[5] A. E. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3D scenes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=765655

[6] R. Szeliski, “mage Alignment and Stitching: A Tutorial,” Microsoft Research, Tech. Rep. MSR-TR-2004-92, Dec. 2006.

[7] D. Gibbins, “3D Target Recognition Using 3-Dimensional SIFT or Curvature Key-points and Local SPIN Descriptors,” in *Defence Applications of Signal Processing 2009 (DASP’09)*, Kauai (Hawaii), 2009. [Online]. Available: http://www.adelaide.edu.au/directory/danny.gibbins

[8] T. Lo and J. P. Siebert, “Local feature extraction and matching on range images: 2.5D SIFT,” *Computer Vision and Image Understanding*, vol. 113, no. 12, pp. 1235–1250, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2009.06.005

[9] F. Tombari, S. Salti, and L. Di Stefano, “Unique Signatures of Histograms for Local Surface Description,” in *ECCV 2010 Part III*, 2010, pp. 347–360.

[10] D. Lowe, “Distinctive image features from scale-invariant keypoints,” 2003. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.8899

[11] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded Up Robust Features,” in *9th European Conference on Computer Vision*, Graz Austria, May 2006.

[12] T.-W. R. Lo and J. Siebert, “SIFT keypoint descriptors for range image analysis,” *Methodology*, vol. 2008, no. 3, pp. 1–17, 2008. [Online]. Available: http://eprints.gla.ac.uk/14123/

[13] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape context,” *IEEE Trans Patt Anal Mach Intell*, vol. 24, no. 4, pp. 509–522, 2002.

[14] R. Bro, E. Acar, and T. Kolda, “Resolving the Sign Ambiguity in the Singular Value Decomposition,” 2007.

[15] F. Tombari, S. Salti, and L. Di Stefano, “Unique Shape Context for 3D Data Description,” in *3DOR’10*, 2010.

[16] H. Hoppe, T. DeRose, T. Duchamp, J. Mcdonald, and W. Stuetzle, “Surface reconstruction from unorganized points,” in *COMPUTER GRAPHICS (SIGGRAPH ’92 PROCEEDINGS)*, 1992, pp. 71–78.

[17] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, “An optimal algorithm for approximate nearest neighbor searching fixed dimensions,” *Journal of the ACM*, vol. 45, no. 6, pp. 891–923, 1998. [Online]. Available: http://portal.acm.org/citation.cfm?doid=293347.293348

[18] P. J. Besl and H. D. McKay, “A method for registration of 3-D shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=121791