Purely Geometric Scene Association and Retrieval
- A case for macro-scale 3D geometry

Rahul Sawhney¹,², Fuxin Li², Henrik I. Christensen³ and Charles L. Isbell¹

Abstract—We address the problems of measuring geometric similarity between 3D scenes, represented through point clouds or range data frames, and associating them. Our approach leverages macro-scale 3D structural geometry - the relative configuration of arbitrary surfaces and relationships among structures that are potentially far apart. We express such discriminative information in a viewpoint-invariant feature space. These are subsequently encoded in a frame-level signature that can be utilized to measure geometric similarity. Such a characterization is robust to noise, incomplete and partially overlapping data besides viewpoint changes. We show how it can be employed to select a diverse set of data frames which have structurally similar content, and how to validate whether views with similar geometric content are from the same scene. The problem is formulated as one of general purpose retrieval from an unannotated, spatio-temporally unordered database. Empirical analysis indicates that the presented approach thoroughly outperforms baselines on depth / range data. Its depth-only performance is competitive with state-of-the-art approaches with RGB or RGB-D inputs, including ones based on deep learning. Experiments show retrieval performance to hold up well with much sparser databases, which is indicative of the approach’s robustness. The approach generalized well - it did not require dataset specific training, and scaled up in our experiments. Finally, we also demonstrate how geometrically diverse selection of views can result in richer 3D reconstructions.

I. INTRODUCTION

The problems of computing similarity and establishing association between range images and/or 3D point clouds of scenes (observed from a viewpoint, henceforth referred to as scene-views) is fundamental to robotics and computational perception in general. It plays an important role in a multitude of applications. Loop closure (identifying a place visited earlier in the trajectory) is intrinsic to metric SLAM ([9]). Localizing with respect to a previously reconstructed map or scene model, and relocalizing after a tracking failure (determining sensor pose without pose priors from trajectory) - both are essential for mapping in practice as well. Different forms of the problem are also key to many navigation scenarios, and in perception tasks such as scene-guided search / foraging or location-based context and activity recognition.

In a minimally restrictive setting, the aforementioned problems (and several others) can be formulated as a retrieval problem - to recognize / identify a scene-view by linking it to stored ones in an assorted, unorganized database. Such a setting would not require any pose priors, spatio-temporal contiguity of collected data¹ or other additional information such as annotations or reconstructed models, and would remove the need to learn a specific pose estimator / regressor for each workspace.

While a lot of progress has been made over the years, including in the retrieval domain, competitive scene association approaches in literature have mostly been reliant on (discriminative) appearance information. Relatively few methodologies work well on noisy, imperfect 3D point clouds or depth images from the real world. Often they critically rely on additional pieces of information available in their target scenario - to prune the association hypothesis space, or obtain strong indirect priors on scene similarity, or enable construction of aggregated spatial information structures to allow its estimation (for instance, [25, 11, 7, 32, 6, 3]) ². Approaches also often operate under limited changes in viewpoint and / or on specific types of scene geometry (such as [42, 11, 29]) or they solve a simplified 2D problem (such as [19]). Understandably, methodologies like above are either use case limited or restrictive. Note that approaches like [10] do not ascertain association at all - these directly solve for 3D poses between pre-associated set of data frames.

The dearth of purely 3D geometric scene association approaches in the real world can be primarily attributed to the considerably more ambiguous and challenging depth / range sensing modality. In general, the modality has high local ambiguity and may not be lavished with information on the whole (in contrast to rgb). Data acquired from commodity 3D range / depth sensing hardware tends to be particularly noisy as well, has several imperfections. Locally smooth, isomorphic and self-similar nature of typical 3D data from indoor or structural environments makes the problem more difficult. Changes in viewpoint, occlusions and partially overlapping views / content significantly exacerbate the problem further.

We present a minimally restrictive retrieval methodology. Our approach affords means to evaluate geometric content similarity between 3D point sets and associate them. We show how it can be utilized to affect geometric diversity as well.

We generate descriptive frame-level signatures directly from range images / point clouds (any additional information or assumptions touched upon earlier are not utilized). We make use of macro scale geometry — 3D geometrical interactions (derived from relative angles and distances) over

¹ Spatio-temporally unordered databases can store data acquired from multiple sensors, at multiple times and from disparate locations; could just constitute of snapshots covering scenes of interest.

² Quite commonly, approaches rely on spatio-temporal contiguity of frames to to obtain priors or accumulate data structures to ascertain the association.

¹ Institute of Robotics & Intelligent Machines, College of Computing, Georgia Institute of Technology, Atlanta, USA raawhney3@gatech.edu, isbell18cc@gatech.edu ² School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, USA lif@engr.orst.edu ³ Institute of Contextual Robotics, University of California, San Diego, USA hchristensen@ucsd.edu
were quite encouraging as well. They not only indicated a significant increase in viewpoint diversity of the retrieved views undergo diversification and subsequent validation. The point clouds are color mapped according to the surface normals - the RGB color of a 3D point is proportional to the component values of its normal.

an arbitrarily large span, between arbitrary surfaces, primitives and structures, and their spatial arrangement (for example between walls, floor and furniture, between fixtures and equipment, or just between various parts of a given structural entity). Such interactions when considered collectively are highly discriminative. They are expressed in a learnt viewpoint invariant feature space (III-A). To characterize a scene-view, high order gradient statistics from a dense set of projected interactions are utilized (Fisher Vector, III-B). To identify a geometrically diverse subset from set of similar retrieved views (III-C), we model a Determinantal Point Process (DPP, III-D). And to establish association with some of the retrieved views, we employ a fine-grained spatial validation scheme which ascertains consistency of rigid geometry overlap (III-E).

The proposed approach not only outperformed the range / depth data baseline, but was also comparable or better than state-of-art RGB and RGB-D approaches (including ones based on CNN 3) - and without relying on any additional pose annotations, apriori reconstructed 3D world models, or assumptions such as spatio-temporal contiguity of training data used by other approaches.

Experiments also indicated the learning to be general - unlike most other approaches, it did not require dataset specific training; a single learnt model performed well across the board. Experiments also indicated the performance holding up under significantly sparser databases, and under significantly increased database scale and diversity. Our empirical evaluations quantifying geometric diversity of retrievals were quite encouraging as well. They not only indicated a significant increase in viewpoint diversity of the retrieved set, but also suggested the efficacy of the proposed approach for richer reconstruction and increased workspace coverage - promising hitherto unexplored application scenarios, such as assistive structural search.

II. RELATED WORK

We refer to only more recent 3D literature among the vast and varied landscape. State-of-the-art loop closure, camera relocalization and place recognition approaches have been primarily based on visual information ([28] presents a recent survey). Many rely on landmark-based features, such as SIFT or ORB, for instance [34, 1, 31, 26]. Approaches such as [38] have focused on the classification problem - one of categorizing similar scenes. [38] utilizes user annotated 3D data to categorize scenes with viewpoint invariance.

Recent state-of-the-art sensor relocalization approaches in real world structural settings [46, 4, 26, 47, 22] are appearance-reliant as well. They also have other critical requirements like scene-specific learning, and / or workspace models or apriori constructed feature clouds (Section IV).

As discussed earlier, high-performing scene association approaches operating solely on 3D range/depth data have been relatively scarce. A significant amount of efforts have been put on local 3D point features, such as [44, 37, 42]. There have also been work based on complete point clouds include variants of Iterative Closest Point, Normal Distributions Transform and aggregated 3D features (often position based, such as height above ground, [13]) over densely sampled keypoints. While they work well under some conditions, their performance deteriorates quickly with increasing change in viewpoint and sensor rotations - [15, 14, 39] amongst others, have noted this as well.

[39] matches surface patches between views operating on range / depth data. Our geometric property extraction is along similar lines, and our validation scheme builds upon it. In contrast to [39], which presents a localized surface patch matching algorithm based on aligning geometric sequences defined over neighborhood patches, this work focuses on capturing holistic scene level content for ascertaining geometric content similarity and retrieval.

A number of successful methods exist for shape-based retrieval. [27] presents a recent survey. Shape retrieval approaches are designed to work with CAD object models or clutter free, object-centric data, often with 3D figure-ground information (in contrast to raw, egocentric scene data from noisy sensors) 4. There have been some successful approaches

3 Note that our approach considers surface patches as far as half a frame apart from the outset – a distinct difference from popular convolutional network based learning approaches that start by building local features.

4 These also involve specific assumptions - for example, watertight manifolds, surfaces with geometric texture, or disparate / distinctive topology.
for 3D object instance detection in clutter, by employing pre-ascertained 3D object templates, for example [21, 2]. More recently, approaches such as [36] have learnt object point clouds to identify 3D shapes with distinct topology.

III. METHODOLOGY

Given a queried scene-view, $V_Q$, and an extant database, $D$, of various views from various scenes, $\{V_n\}_D$ - our algorithm

(a) Retrieves a set of views which have structurally similar content as $V_Q$, b) Identifies a geometrically diverse subset of views from this retrieved set, and c) Ascertains whether some of these views pertain to the same scene as $V_Q$ (Figure 1).

We denote $\{x_i\}_{i=1}^X$, $x_i \in X$ to indicate $V_X$’s segmentation into smooth surface patches. $\{X_h\}_{h=1}^H$ denotes the segmentation hierarchy then. Hierarchy generation is outlined in Section III-F. To simplify notation, we only indicate the hierarchy level $h$ when it improves clarity.

A. Geometric feature space description

Geometric property set extraction: For a given view $V_X$, at a particular segmentation level - we first express each patch $x_i$ through a 13-D vector set, $F'_{x_i}$ of robust, viewpoint agnostic and macro scale 3D geometric properties. These are derived by utilizing 3D relationships relative to other patches in $x_i$’s neighborhood, $N_x$ (along similar lines as [39]). Note that $N_x$ is large, non-local - it could span the entire segmentation, $X - x_i$. Neighboring patch count, $|N_x|$, is indicated as $c_{x_i}$. For a patch $x_i \equiv \mu \in X$, we denote its mean surface normal as $\hat{n}_\mu$ and its 3D mean as $l_\mu$. Denoting $\alpha$ to indicate a patch in $\mu$’s neighborhood, with $\hat{n}_\alpha$, $l_\alpha$ denoting its normal and mean respectively - an orthonormal basis can be derived from the spanning vectors $\hat{n}_\mu$ and $r^\alpha_\mu = l_\alpha - l_\mu$ through the Gram-Schmidt process. Figure 2 illustrates this. It also formulates the resultant orthonormal basis, $<\hat{u}^\alpha_\mu, \hat{v}^\alpha_\mu, \hat{w}^\alpha_\mu>$, where $\hat{u}^\alpha_\mu$ is the unit vector in the direction of $r^\alpha_\mu$. Note that coordinate frame spanned by this orthonormal basis is agnostic (invariant) of the sensing viewpoint, since it is a reference frame local to $\mu \equiv \alpha$. Also note that this basis is seldom degenerate, as $\hat{n}_\mu$ and $r^\alpha_\mu$ are rarely colinear, especially when data frames are captured from a projective sensing process.

For each neighboring surface patch $\alpha$ in $\mu$’s neighborhood, $\mathcal{N}_\mu$, we are able to thus extract the following vector of viewpoint invariant properties, $(f^\alpha_\mu)_{\forall \alpha \in \mathcal{N}_\mu}$:

$$f^\alpha_\mu = [\theta_{\hat{n}_{\alpha}, \hat{n}_\mu}, \theta_{\hat{u}^\alpha_\mu, \hat{n}_\mu}, \theta_{\hat{v}^\alpha_\mu, \hat{n}_\mu}, v^\alpha_\mu \cdot \hat{n}_\mu, \hat{n}_\alpha \cdot \hat{u}^\alpha_\mu, \hat{n}_\alpha \cdot \hat{v}^\alpha_\mu, \hat{n}_\alpha \cdot \hat{w}^\alpha_\mu, \ldots, r^\alpha_\mu \cdot (\hat{n}_\alpha \times \hat{n}_\mu), ||\hat{u}^\alpha_\mu||, ||\hat{v}^\alpha_\mu||, \text{sgn}(\hat{n}_\alpha \cdot \hat{u}^\alpha_\mu), ||r^\alpha_\mu|| \cdot \text{sgn}(\hat{n}_\alpha \cdot \hat{v}^\alpha_\mu), \ldots, ||\hat{v}^\alpha_\mu|| \cdot \text{sgn}(\hat{n}_\alpha \cdot \hat{w}^\alpha_\mu), ||\hat{w}^\alpha_\mu|| \cdot \text{sgn}(\hat{n}_\alpha \cdot \hat{w}^\alpha_\mu), ||\hat{v}^\alpha_\mu|| \cdot \text{sgn}(\hat{n}_\alpha \cdot \hat{w}^\alpha_\mu), ]^T$$ (1)

The $\theta$ above refers to the angle between the indicated vectors and $\times$ represents an outer product. $\text{sgn}(.)$ is a robust signum function that clamps to zero when its parameter $\notin [\cos^{-1}(P_1 - \epsilon_\theta), \cos^{-1}(\epsilon_\theta)]$, with $\epsilon_\theta$ accounting for allowable tolerance to angular noise.

The feature vector $f^\alpha_\mu$ basically represents an overcomplete characterization of relative properties between the two patches - formulated in a viewpoint agnostic fashion. The first part (first 9 features) captures angular relationships between $r^\alpha_\mu$, $\hat{n}_\alpha$ and $\hat{n}_\mu$, characterizes $\hat{n}_\alpha$ in the invariant frame derived from $r^\alpha_\mu$ and $\hat{n}_\mu$, and characterizes $r^\alpha_\mu$. The second part (remaining 4) consists of robustified features - as a measure against noises arising due to estimation from real world, noisy data. Signs of projected normals’ components are captured through robust signum functions and are augmented with the magnitude of relative displacement vector.

Feature space projection: A patch’s property set $F'_\alpha = \{f^\alpha_\mu\}_{\alpha \in N_{x_i}}$ is then projected onto a subspace learnt through nonlinear independent component analysis ([23]). The projection reduces redundancy in $f^\alpha_\mu$, denoising and making the components more independent. Importantly, this fits with the component independence assumption made in Section III-B to train Gaussian mixtures with diagonal covariances.

The feature space projection results in a 12-D feature vector set $F_{x_i} = \{F'_\alpha\}_{\alpha \in N_{x_i}}$. By considering the patches in $x_i$’s large scale neighborhood, $N_{x_i}$, the feature set $F_{x_i}$ can thus robustly express the 3D geometry in $x_i$’s non-local neighborhood. An aggregation of such feature sets arising from all the patches, $F^X = \{F_{x_i}\}_{i=1}^X \equiv \{F'_\alpha\}_{x_i \in X, \alpha \in N_{x_i}}$, can thus invariantly and richly express the geometry of the entire scene as captured by $V_X$. Finally, the above procedure is repeated for each level in the segmentation hierarchy, to capture fine as well as coarse details. This results in a hierarchy of aggregate feature vector sets, $\{F^h\}_{h=1}$. 

Fig. 2: For a given patch $\mu$, relative and invariant 3D properties are extracted with respect to patches in a non-local neighborhood. To facilitate that, an orthonormal, viewpoint agnostic frame is derived using the Gram-Schmidt process.

Fig. 3: Geometric properties are extracted over a hierarchy of patch segmentations. At each segmentation level, the aggregate sets of properties is first mapped to a viewpoint invariant geometric feature space. These are then jointly encoded as a view level signature using fisher vector embedding.
B. Encoding feature space statistics

To obtain a descriptive signature for a given view, \( V_X \), we encode the aggregated feature sets using Fisher vector embedding (FV, [20, 35]) - this captures the normalized gradient of the log-likelihood of the feature sets. The Fisher kernel theory, first presented in [20], introduces a similarity kernel, arising as a consequence of maximizing the log-likelihood of generatively modeled data. In this paper, Gaussian Mixture Models (GMM) were used to model the feature space distribution.

Given a learnt GMM, \( P_\theta \), parameterized as \( \theta=\{p_y, \nu_y, \Lambda_y\} \), the FV embedding of the aggregate feature set \( F_X \), indicated as \( \phi(F_X) \), is obtained as \( \phi(F_X)=L_\theta \nabla \log P_\theta(F_X) \). Here, \( L_\theta \) is the Cholesky decomposition factor of the inverse Fisher Information Matrix, and \( \nabla \log P_\theta(F_X) \) is the score function (log-likelihood gradient). Following similar analysis as [35], under assumptions of diagonal covariance matrices, \( \Lambda_y \), and independence of the samples, \( f_{x_i}^g \), the embedding evaluates as \( \phi(F_X)=m_0^g m_1^g \ldots m_g^g m_0^g \nu_g^T, m_0^g \nu_g \ldots m_G^g )^T \) - where \( m_g \) and \( \nu_g \) respectively capture the normalized zeroth, first and second order statistics of the sample set, that falls in the \( g \)-th mixture component of the GMM. \( \phi(F_X) \) has a dimensionality of \( d_\theta = (2d_F + 1) \cdot G \), where \( G \) is number of mixture components, and \( d_F = 12 \) is the dimensionality of our geometric feature space. Below, 1 denotes an all-one vector and \( \pi_{ij,g} \) is the joint pdf of \( X \) conditioned on \( F_X \).

\[
\begin{align}
\pi_{ij,g} &= \frac{\exp\left[-\frac{1}{2}(f_{x_i}^g - \nu_g)^T \Lambda_g^{-1}(f_{x_i}^g - \nu_g)\right]}{\sum_{j=1}^{c_G} \exp\left[-\frac{1}{2}(f_{x_i}^g - \nu_g)^T \Lambda_g^{-1}(f_{x_i}^g - \nu_g)\right]}.
\end{align}
\]

\[
\begin{align}
m_{0}^g &= \frac{1}{c_X \sigma_g \sqrt{p_y}} \sum_{i=1}^{c_X} \sum_{j=1}^{c_G} \pi_{ij,g} - p_y \tag{2a}
m_{1}^g &= \frac{1}{c_X \sigma_g \sqrt{p_y}} \sum_{i=1}^{c_X} \sum_{j=1}^{c_G} \pi_{ij,g} \Lambda_g^{-1/2}(f_{x_i}^g - \nu_g) \tag{2b}
m_{2}^g &= \frac{1}{c_X \sigma_g \sqrt{2p_y}} \sum_{i=1}^{c_X} \sum_{j=1}^{c_G} \pi_{ij,g} \left[ \Lambda_g^{-1}(f_{x_i}^g - \nu_g) (f_{x_i}^g - \nu_g)^T - I \right] \tag{2c}
\end{align}
\]

\( \phi(F_X) \) is then component-wise square root normalized (by replacing each component, \( a' \) of \( \phi(F_X) \) by \( |a'|^{1/2} \text{sign}(a') \)), and \( \ell_2 \) normalized. The square root normalization serves to alleviate the dominant effect of relatively indiscriminate samples occurring with high frequency (for example, arising from patches on a wall or ceiling) and the \( \ell_2 \) normalization helps generalization across different scenes by normalizing the energy content. The desired view signature vector for \( V_X \), denoted as \( \psi(X) \), is obtained by evaluating the embedding at each level in hierarchy, and concatenating them - \( \psi(X) = [\phi(F_X)^T, ..., \phi(F_X)^T] \).

C. Similarity and Retrieval

The thus obtained view signature, \( \psi(X) \), captures discriminative 3D geometrical properties, and is robust to viewpoint changes, sensor noise, occlusions and other data imperfections by design. As experiments indicate, a metric based on such view signatures is a reliable measure of 3D geometric similarity. We tried \( \ell_1 \) & \( \ell_2 \) distance metrics, and used \( \ell_1 \) for all experiments in the paper as it performed better. Thus the similarity between two given views \( V_X \) & \( V_Y \) can be denoted as, \( s(X,Y) = -\sum_{i=1}^{2G} \left| \phi(X) - \phi(Y) \right| \).

Given a a queried view, \( V_Q \), and a database \( D \) of view signatures, one can thus retrieve a set of putative view associations in the geometric sense through nearest neighbor queries. We indicate this retrieved set of putatively associated views as \( R = \{ V_X \}_{X=1}^{C_W} \).

D. Diversity Sampling with Determinantal Point Processes

Depending on the distribution of scenes’ views in the database, \( R = \{ V_X \}_{X=1}^{C_W} \) could be overwhelmed with views which are near duplicates (all being very similar to each other, hence almost equally similar to the queried view). This may not be desirable since the subset of top retrievals could just be flooded with near duplicates of false putative associations, resulting in complete failure. By filtering out near duplicates, a diversity based subset selection procedure may still be able to salvage correct, albeit lower ranked, putative associations present in \( R \) with further post-processing validation.

A diverse set of retrievals is generally desirable. It would provide assisted with possibly complementary information, which could be made use of thereon. For instance, it could be potentially beneficial in reconstruction or coverage tasks, where diverse viewpoints observing the environment with only partially overlapping content are more desirable than having redundant views from nearly the same perspective. A querying human user could also be better assisted by being provided with a diverse set of the retrievals to choose from.

Determinantal point processes ([24]) are employed to select a diverse subset of candidate views, \( c \), from \( R \). A point process \( P_L \) is called an \( L \) - ensemble \( k \)-determinantal point process if for every random subset, \( c \), of \( R \), such that \( |c| = k \), drawn according to \( P_L \), we have \( P_L(c, R) = \frac{\det(L_c)}{\det(L_R)} \). Here is a symmetric positive semi-definite similarity matrix indexed by the elements of \( R \). \( L_C \) is the principal minor (submatrix) with rows and columns from \( L \) indexed by the elements in subset \( c \). Thus the probability of selecting a subset \( c \), \( (|c| = k = c_T) \) elements is directly proportional to the determinant of the submatrix indexed by it. Note that higher diagonal values would proportionately encourage their inclusion in a selected subset \( c \) as they lead to higher determinants. Similarly, the off-diagonal values determine correlation between different elements, and a high value decreases the determinant overall. Thus two elements with a high similarity value tend not to co-occur in \( C \). DPP sample sets are therefore able to balance the net significance of their constituent elements with their diversity. We modeled \( L \) accordingly as follows

\[
\{L\}_{X,Y} = \rho_X \rho_Y \kappa s(X,Y)^\omega, \quad 1 \leq X, Y \leq C_R
\]

where \( \rho_X = e^{1 \frac{s(X, Q)}{\sigma}}, \quad 3X \in R \) models the similarity of a retrieved view \( V_X \) to the queried view \( V_Q \). The similarity between two given views \( V_X \) and \( V_Y \) is captured by the rightmost term. Positive valued parameters \( \sigma, \omega \) and \( \kappa \) can be tuned to balance the need for both diversity and similarity to \( V_Q \). A lower sigma would induce a higher resolution in
similarity scores between retrieved views, and hence would result in a more diverse subset selection.

While the MAP inference on $\mathcal{P}_L$ to determine the most probable subset is NP-hard, efficient sampling algorithms exist which provide good approximate solutions in practice. For our purposes, a greedy procedure based on [24] which results in $O(k \log k)$-approximation worked well.

E. Validating candidate views for association

We employ a finer grained spatial validation step before finally associating the queried view with some of the views in the candidate set, $\mathcal{C} = \{V_X\}_{X=1}^C$. This is done by directly leveraging the rigid 3D spatial arrangement of surface patches to ascertain surface alignment. We make use of the patch matching scheme presented in our prior work [39]. It utilizes a sequence alignment scheme over similarly motivated patch properties to find standalone correspondences based on 3D neighborhood similarity. A semi-dense set of correspondences can be ascertained. Rigid transform between two views of a given scene can then be robustly, accurately computed through consensus of patch associations.

When views from scenes with different geometrical content are matched through [39], the matches would likely be inconsistent with respect to the computed transform. We exploit this understanding to validate associations with candidate views. For each candidate view, $V_X \in \mathcal{C}$, and the queried view, $V_Q$, we utilize randomly sampled patches to estimate rigid transformations both ways, that is, $T^Q_X = (R^Q_X, t^Q_X)$ and $T^X_Q = (R^X_Q, t^X_Q)$ and check whether they are consistent with each other. We ascertain a candidate $V_X \in \mathcal{C}$ as associated with $V_Q$ when $\|\log(R^Q_X R^X_Q)\|_2 \leq \epsilon_{val}$, $\|t^Q_X + t^X_Q\|_2 \leq \epsilon_{val}$ - we are basically ensuring that the magnitude of the rotation and translation components in the residual transform, $T^Q_X T^X_Q$, are below certain thresholds $\{\epsilon_{val}, \epsilon_{val}\}$. Parallelism of this step is facilitated by the nature of the linear transformations, so a straightforward fashion. The computational bottleneck arises during validation, which is quadratic in number of superpixels (~1s for segmentation with 1000 superpixels at finest level, but again naturally parallelizable). Note that it suffices to validate at a coarse hierarchical level (~ 250 superpixels) — the result, $T^X_Q$, can then be used as reliable initialization and be quickly refined iteratively as per task.

F. Further details and discussion

The approach is amenable to any boundary-preserving patch segmentation scheme, as long as it results in superpixels / patches that are geometrically regularized for smoothness and compactness. For example [33], which segments volumetrically, could be used while working with point clouds; and surface segmentation schemes such as one presented in [39] could be employed when working with depth / range images. Both [39] and [33] performed well in our experiments. Starting with the base level segmentation, a segmentation hierarchy can be built in either fine to coarse (agglomerative), or coarse to fine (divisive) fashion. Each successive level has patches reduced (increased, in case of divisive) by a constant factor - this can be approximately ensured by employing K-Means in 3D with near uniform surface component seeding (any resultant patches below a certain size / surface area are merged back). We used four levels of segmentation hierarchy ($H = 4$). The number of mixture components were also kept fixed, $G = 1250$. The GMMs were learnt through an expectation maximization scheme, and the mixture components were initialized from the result an iteration of K-Means++ procedure. Our empirical analysis indicated the learnt feature space distribution to be general for similar sensor types 5. Figure 4 suggests that as well. In fact, a single set of Gaussian mixture (and ICA) models were utilized for all the experiments shown in the article (except Figure 4).

In practice, for efficiency, while encoding feature space, we suffices to approximately ascertain $F^X_h$ by sampling patches from $X_h$, and subsequently sampling the neighborhoods of the sampled patches. This partly corroborates our assertion that the methodology is robust to occlusions. Databases were indexed as KD-trees. Our current straight up implementation is not optimized for efficiency (on a 4.2 GHz, 4 core setup, III-A - III-D takes ~ 3 ms, 1000 superpixels), though the methodology is GPU parallelizable. Most of the procedures outlined in Sections III-A, III-B, III-C, III-D and III-E can be GPU parallelized in a straightforward fashion. The computational bottleneck arises during validation, which is quadratic in number of superpixels (~1s for segmentation with 1000 superpixels at finest level, but again naturally parallelizable). Note that it suffices to validate at a coarse hierarchical level (~ 250 superpixels) — the result, $T^X_Q$, can then be used as reliable initialization and be quickly refined iteratively as per task.

IV. Experiments

In all experiments, the method indicated ‘R’ refers to our retrieval approach (till Section III-B), without the diverse subset selection and validation steps. ‘DR’ refers to our approach till Section III-D, with diversification but without the validation step. ‘VDR’ would then refer to the complete approach, resulting in the set $C_{vld}^r$ - diverse retrievals which have been validated through rigid overlap consistency. The retrievals in both the sets $C$ and $C_{vld}$ follow the same order (by $s(X, Y)$) as they appear in the initial retrieval set $R$. All analysis is done on the top few results from these sets.

The retrieval and association problems can be subjective - two views with only partially overlapping geometric content can be evaluated differently by users. We employed an objective measure - evaluating our retrieval approach on a sensor relocalization task. We utilized the 7-scenes datasets from [12, 40], the standard benchmark for indoor RGB/RGB-D relocalization. The objective is to localize the sensor (ascertain pose) with respect to the workspace within the maximal allowable translation and orientation errors (5 cm and 5 deg respectively). The datasets are collected from different workspaces (although some scenes in RedKitchen and Pumpkin

\[ Sufficient number of GMM components should be utilized to span the extent of the geometric feature space. This is a function of maximum scene scale captured, and thus sensor range.\]
As opposed to mapping, visual odometry or semantic scene datasets such as [43, 17, 48, 8, 18] these either do not have enough loop closures and/or are synthetic, or lack ground truth for quantitative evaluation or standard train-test splits for loop closure.

6 As opposed to mapping, visual odometry or semantic scene datasets such as [43, 17, 48, 8, 18] these either do not have enough loop closures and/or are synthetic, or lack ground truth for quantitative evaluation or standard train-test splits for loop closure.

Table I: The presented approaches (R, VDR) are compared with baselines through localization accuracies on the standard 7-scenes datasets from [12, 40]. All methods utilize RGB-D data during training, except [40] D, and our R and VDR, which are based on range / depth data. During test time, the three leftmost approaches only take RGB images as input, while the three rightmost approaches only take range / depth images - the rest operate on RGB-D. Average indicates the average among the 7 datasets. Combine indicates performance when jointly considering all 7 scenes as a single database. VDR outperforms all the RGB-D approaches while using depth information only. R performs very well as well, outperforming all but two RGB-D approaches. As opposed to mapping, visual odometry or semantic scene datasets such as [43, 17, 48, 8, 18] these either do not have enough loop closures and/or are synthetic, or lack ground truth for quantitative evaluation or standard train-test splits for loop closure.

As Table I shows, VDR achieved state-of-the-art results through pure geometry alone - without needing any additional annotations, assumptions or appearance features. Equally promising were the results from R which were obtained by simply using the first retrieval in R (no diversification or validation), which were better than all baselines but two. DR gives the same results as R in the relocalization experiments and is hence not shown. This is because the accuracies were evaluated with respect to only the top retrieval - this is the same for R and DR since the greedy algorithm we used for k-DPP automatically selects the top-scoring retrieval as the first one. These results support our hypothesis that macro-level 3D geometry holds immense discriminative information.

In the last row of Table I, we combined all training data from the 7 datasets into one single database, and evaluated accuracies of the combined test splits. As can be seen, the results held up quite well in the combined experiment, when the database size and complexity (variety, aliasing) was drastically increased.

We also tabulated the affect of significantly reducing the database sizes - by re-evaluating results with databases built from only 1/15th and 1/20th of the available train-
Fig. 6: Quantifying diversity. Left, Middle: The average relative translation of the retrieved views with respect to the queried view. One can see DR improves diversity over R, and VDR improves over VR. Right: Efficacy of diverse viewpoints for reconstruction task. The average number of voxels (in a 8 cm$^3$ occupancy grid) occupied by ground truth reconstructions from the first five validated retrievals from VR and VDR are plotted. From the same number of initial views, VDR results in richer reconstructs that capture significantly more voxels in the scene.

splits for each dataset. With a much sparser coverage of the environment, both retrieval and subsequent validation and localization becomes much more difficult. The frames were sampled at uniform intervals, thus may have steep viewpoint changes, much reduced content overlap and significantly increased occlusions. As Figure 5 indicates, the accuracies of both $R$ and VDR held up quite well. This is indicative of the approach’s robustness to these practical challenges.

The approach generalizes well. Our experiments do not suggest a need for scene specific training - a single set of learnt gaussian mixtures and ICA projection matrices were utilized in all our experiments (except Figure 4). The training data was taken from the train split of Redkitchen in [40], and from datasets in [48, 43], a reasonably rich and diverse set of samples. Figure 4 shows the robustness of the GMM parameters with respect to the dataset used to train it. As can be seen the results stay consistent.

Finally, we conducted experiments to quantify the effect of our diversification approach, and its role in generating significantly richer reconstructions. As Figure 6(left, middle) show, the diversity of retrieved viewpoints is greatly improved due to our DPP-based diversification. Note that DR and VDR select views which are not only further off than the queries (higher relative mean), but result in view sets which have significantly more viewpoint variance amongst themselves as well (significantly higher standard deviations). And as Figure 6(right) shows, the reconstruction volume improves significantly when a diversified set of views is utilized. Figure 7 shows a qualitative example. One can see that the diversified retrievals are significantly more diverse, from varied viewpoints, and are resulting in an appreciably richer reconstruction. In general, retrieval as well as geometric diversity is often desirable - apart from reconstruction, it would prove useful in other tasks such as structure and semantic analysis, and 'human in loop' selection tasks.

V. CONCLUSION

We presented a robust solution to the problems of measuring 3D geometric similarity between 3D range images or point clouds, and determining whether they come from the same scene. A general-purpose retrieval approach was proposed, based on encoding (FV) of viewpoint-invariant features that are hand-crafted to capture 3D geometry at macro scales. The approach performed well in real world settings - including ones that involved sharp viewpoint changes, partially overlapping and occluded content. It scaled well, and did not require scene-specific training - making it useful in a variety of scenarios. As experiments established, the approach is powerful and did better than specifically fitted solutions such as CNNs trained on RGB or RGB-D data. Furthermore, we introduced a way to obtain geometrically diverse retrievals (DPP), and showed how such retrievals can help generate richer reconstructions. Interestingly, in contrast to CNN approaches which begin with a local neighborhood, our approach utilized macro scale features from start. The combination of both paradigms would be explored in future work, for this and other tasks involving 3D recognition.

Acknowledgements
RS and FL are supported in part by NSF grant IIS-1464371.

REFERENCES


S. Li and A. Calway. RGBD relocalisation using pairwise geometry and concise key point sets. In Robotics and Automation (ICRA), 2015.


