Towards 3D VR-Sketch to 3D Shape Retrieval

Ling Luo1*,*2 Yulia Gryaditskaya1*,*2 Yongxin Yang1*,*2 Tao Xiang1*,*2 Yi-Zhe Song1*,*2

Abstract

Growing free online 3D shapes collections dictated re- search on 3D retrieval. Active debate has however been had on (i) what the best input modality is to trigger retrieval, and (ii) the ultimate usage scenario for such retrieval. In this paper, we offer a different perspective towards answering these questions – we study the use of 3D sketches as an input modality and advocate a VR-scenario where retrieval is conducted. Thus, the ultimate vision is that users can freely retrieve a 3D model by air-doodling in a VR environment. As a first stab at this new 3D VR-sketch to 3D shape retrieval problem, we make four contributions. First, we code a VR utility to collect 3D VR-sketches and conduct retrieval. Second, we collect the first set of 167 3D VR- sketches on two shape categories from ModelNet. Third, we propose a novel approach to generate a synthetic dataset of human-like 3D sketches of different abstract levels to train deep networks. At last, we compare the common multi-view and volumetric approaches: We show that, in contrast to 3D shape to 3D shape retrieval, volumetric point-based approaches exhibit superior performance on 3D sketch to 3D shape retrieval due to the sparse and abstract nature of 3D VR-sketches. We believe these contributions will collectively serve as enablers for future attempts at this problem.

1. Introduction

3D model retrieval has become an important topic due to a growing number of free online 3D repositories. It finds applications in CAD design, 3D printing, 3D animation and movies production, where the time required to model a 3D shape can be strongly reduced by relying on retrieval from existing 3D shape collections. Various in- put modalities have been tried – images and rough 2D sketch at first, with latest research focusing on 3D to 3D, i.e., using 3D scans or existing 3D shapes.

Despite great strides made, two salient questions still re- main (i) what constitutes the best input modality to initiate 3D retrieval, and (ii) under what usage scenario can such retrieval be best facilitated. For the former, 2D sketches and images are both in 2D, therefore can not reach the level of details desired, and the 2D-3D domain gap can also be counter-intuitive. The 3D-based paradigm on the other hand dictates existing 3D models to be readily available, which to some degree forms a “chicken-and-egg” problem (i.e., where/how to source the input model at the first place). As for the latter, apart from 2D sketches which are freely de- fined by the user, all other usage scenarios do not offer flexibility in terms of the input desired – one can not easily alter an image, not to mention a 3D model/scan.

In this paper, we offer a new perspective on 3D shape retrieval – we advocate the use of 3D sketches as a new input modality. This new 3D-VR sketch to 3D shape paradigm not only enables detailed retrieval as per the common 3D model to 3D model setting, but also simultaneously offers a degree of flexibility found elsewhere in 2D sketch-based retrieval. Our ultimate vision is as follows: with a specific 3D model in mind, one emerges into a VR environment, roughly sketch out the model using handheld controllers, press retrieve and then relevant models would start populating the VR environment.

Our first contribution is therefore coding a VR environment for the said purpose. With this VR environment, we collected the first human VR-sketch dataset. 10 users with no artistic background were hired to produce a total of 167 3D VR-sketches from the two categories from ModelNet: chairs and bathtubs. Some examples are shown in. Since the collection process is both time- and cost- sensitive, we additionally propose the first 3D model to 3D sketch generator, and construct a synthetic dataset of 3D sketches. A key trait of our generator is that it can produce 3D sketches of different abstraction levels, effectively simulating that found in real-human sketches. Through training a series of deep 3D VR-sketch to 3D retrievals model using the real and synthetic datasets, we drive out a few important insights (i) retrieval performance drops with an increasing abstraction level, and (ii) models trained with synthetic sketches can already reach a decent performance level when tested on human sketches.

At last, we experiment with different shape/sketch representations, and state-of-the-art losses commonly used in 3D model retrieval but re-purposed for our problem. We focus on investigating the domain gap between 3D shapes and 3D sketches, due to the two key properties of sketches: sparsity: full 3D models versus sparse sketched lines, and abstraction: 3D models are geometrically perfect, while sketches are subject to deformations. In particular, we examine multi-view versus point-based 3D representations, and show that the later is more robust to both sparsity and an increased level of abstractness. We further propose an architecture with a reconstruction path to explicitly allow for change in abstractness.

In summary, our contributions include: (i) a new perspective on 3D model retrieval, where 3D VR-sketches are used for the first time to conduct retrieval, (ii) a dataset of human 3D VR-sketches, collected using a purpose-built VR environment, (iii) an approach to generate a synthetic 3D sketch with a variable level of abstractness, (iv) comprehensive evaluations using recent 3D shape retrieval models re-purposed to the task of 3D sketch-based retrieval, to drive out insights, plus a novel regularization track to address the sketch-model domain gap.

2. Related work

When dealing with 3D shape retrieval from a single image or a 3D shape, existing work is divided into two large groups based on the shape representation used: view-based or volumetric. The volumetric representation can be further broken down to point-cloud or voxel-based. In this work we show that on the task of 3D sketch to 3D model retrieval point-cloud representation beats multi-view approaches due to better handling of the sparsity of 3D sketches.

2D image- and 3D shape-based retrieval. The retrieval problem in a multi-class setting is closely related to the shape classification problem, where the intermediate embedding of an image or a 3D shape is used for retrieval. A vanilla approach for multi-class shape classification is to use a soft max cross-entropy loss. Others works specifically tackle retrieval , where the triplet loss and its variants have become a common standard. Amongst these, represents the state-of-the- art for 3D model retrieval. It combines triplet and center losses , to solve for their respective drawbacks by using the class center as a positive sample. In this work we evaluate both the triplet and the triplet center losses on the task of 3D sketch-based retrieval and combine them with an additional Reconstruction Loss, tailored to the 3D sketch-based retrieval problem.

Due to a domain gap between images and 3D shapes, or target and query 3D models being from different distributions, Siamese or Heterogeneous network architectures can be more beneficial for a certain problem. In this work, we compare these two types of architectures for our multi-view baseline, and show that due to a strong domain gap between a 3D sketch and a 3D shape, the Heterogeneous architecture by far beats the Siamese one.

3D sketch-based 3D model retrieval. So far there was very little work on 3D sketch-based retrieval, especially in recent years. Most of them work with sketches collected using Microsoft Kinect, which not only has limited tracking accuracy, but the collection interface is also counter-intuitive, where 3D sketching is performed while visualizing 2D projections. As a result, sketches collected mostly have low fidelity and exhibit less details. Our VR-sketches are completely different: (i) visualization and sketching are both performed in 3D, and (ii) the use of the latest VR technology offers a high precision. Together, they ensure our VR-sketches are of high-fidelity and rich in de- tails, thus more fitting for retrieval. A notable exception is the work by Giunchi et al., yet they address a different problem of 3D model retrieval for dense-color VR sketches (and optionally base 3D shapes), while we target at a more simplistic and abstract shape representation – a sparse set of single-color lines. Note that we can not find the said dataset in an open access, thus can not offer a direct comparison. Our NGVNN baseline is however already superior to the state-of-the-art used by and Non-photorealistic rendering (NPR). NPR is an old graphics and vision problem, for a detailed overview of existing methods for generating 2D NPR rendering we refer the interested reader to a recent report. The only at- tempt to produce the representation that resembles a 3D sketch automatically form a 3D shape was proposed by Li et al. as a concatenation of the shape views from six canonical viewpoints. They use this representation to show that the performance of their non-learning method on out- line to 3D model retrieval task significantly outperforms the performance of the 3D sketch to 3D model retrieval. This experiment indicates that such simple shape representation is not sufficient as training data for 3D sketch-based 3D shape retrieval. Our experiments support that: The network trained on sketches with higher abstractness values significantly outperforms the network trained on clean sketches. To the best of our knowledge, we are the first to propose a method to generate abstract 3D sketches from 3D models.

3. Datasets

Collecting a full dataset of 3D human sketches is a labor intensive task. We collect a small dataset of human sketches that we use to guide the synthetic dataset generation. We as well use it as a test set to validate that the network trained on the proposed synthetic data generalizes well to human sketches. To obtain the training data, we reside to a common strategy of generating a synthetic training data.

3.1. Selected shapes

For training and testing our sketch-based retrieval mod- els we use the repaired clean manifold meshes1 from ModelNet10. We use all 10 classes from ModelNet10, with the total of 3958 shapes. We split the dataset into train, validation and test sets, which contain 2847, 317 and 792 shapes, respectively, ensuring a proportional split of shapes of each class between three sets.

3.2. Human 3D sketch dataset

Task. We are targeting 3D sketches created by novices, which can be viewed as an equivalent of the quick 2D sketches from the QuickDraw dataset. To enable the usage of the collected sketches for fine-grained retrieval evaluation we built a dataset of paired sketches and 3D models. We experimented with a setting, where one can observe a 3D model in VR for an unlimited duration of time, and then is asked to sketch from memory. We observed that under this scenario the participants were sometimes omit- ting features important to accurately testing fine-grained retrieval, and instead let the participants to sketch over a reference 3D model. We provide in the supplemental a qualitative evaluation of the sketches from memory and the retrieval performance on such inputs. To mimic the level of detail that can be expected from the 3D sketches from the imagination, we opted to use wide ribbon lines, but do not pose any constraints on sketching style or level of detail.

Participants. We selected 139 chairs and 28 bathtub shapes from ModelNet10’s test set, and hired 10 participants, who have no art background or 3D sketch experience. The shapes were randomly divided into 10 subsets, consisting from 13-14 chairs and 2-3 bathtubs each. Each participant was assigned with one of these subsets. There is no duplicated shapes between subsets.

Interface. Although there are general-purpose VR painting and design software’s that enable users to draw directly in 3D (such as Google’s Tilt brush2 , and Facebook’s Quill3 ), they do not serve our purpose in full: (i) we would like to record detailed stroke-based information, (ii) we require the option of displaying a reference model for data collection, and (iii) we want a fresh code base for any additional functionalities in the future (e.g., sketch-based 3D editing). We implemented our custom 3D sketching environment based on Oculus Rift platform and Unity engine.

3.3 Synthetic training data generation

As a first step towards obtaining a synthetic sketch representation of a 3D shape, we extract detailed curve-networks with the method of Gori et al. This method is de- signed to produce a curve network that preserves well shape details, which are, though, uncommon for human novices sketches We observe that novices not only omit small details, but tend to represent thin volumetric de- tails with single lines and moreover often omit subset of feature lines. To obtain human sketch appearance, we focus on two aspects: level of detail and mechanical inaccuracies. We first perform details filtering and lines consolidation, followed by lines filtering. We then break the long curve chains into shorter strokes, to which we apply a set of local and global transformations .

3.3.1 Curves network extraction.

FlowRep method requires an input to be a curvature- aligned quad-dominant mesh. For processing efficiency, we simplify original triangular meshes by decimation before converting them into quad-dominant meshes using Blender, but the quality of conversion does not always comply with the requirement of FlowRep. Thus, in this stage, we filtered nearly 20% of original dataset which failed to be processed by FlowRep. This can be alleviated in the future by replying on more advanced quad-meshing algorithms.

3.3.2 Details filtering and consolidation.

The code by Gori et al. returns networks of curves in a form of chained edges of an input mesh. We first break these chains into several if the angle between two consequent edges is smaller than 135 degrees. We then remove all the chains those length is smaller than 10% of the smallest of height and width of the bounding box of the original shape. We re-sample all the chains using Ramer- Douglas-Peucker algorithm with an accuracy parameter set to 0.02dmin. Finally, we iteratively go over all the chains and compute for each chain the closest, tangentially aligned chain. If the distance between such two chains is smaller than 5% of the largest length of the two chains, the chains are substituted by their aggregate curve. These steps allow to remove small details and close to each other lines. Yet, the simplified curve networks contain many more lines than can be found in majority of 3D sketches by novices.

3.3.3 Abstraction.

In this section we describe our approach to obtain a 3D sketch with different levels of detail and mechanical inaccuracies, which we jointly refer to as a level of sketch abstractness la. We constraint la to be between 0 and 1.

Level of detail (I) To reduce the number of lines we first compute the similarity matrix using a discrete Frechet distance among chains. When computing the Frechet distance we first align the strokes at one of their end points, by translating one of the strokes. We then perform the grouping according to this similarity matrix with Agglomerative Hierarchical Clustering, where the number of cluster ncluster is a function of la and the number of chains in the sketch nchains : ncluster = max(nchains (1 − 0.8la )/2, 10). (II) We next for each cluster iteratively select a pair of two most distant lines in terms of their absolute positions and compute the mean distance from all the lines dmean in the cluster to these two lines. We remove all the lines in a cluster for which the distance to any of the two selected lines is less than dmean. (III) After getting the reduced set of chains, we break long chains into several shorter chains, which we refer to as strokes. We split each chain at vertices, where the curvatures are twice larger than the mean curvature of an original chain. We further filter out short strokes whose lengths are smaller than 0.2smax to avoid tiny details, where smax is the largest dimension of an input network.

Mechanical inaccuracies To mimic mechanical and perspective inaccuracies of human sketches we apply to each stoke a set of global and local deformations. We first ap- ply a global translation, rotation and scaling, depending on the given level of abstraction la. To achieve this we de- ploy an auxiliary parameter t, which is randomly sampled from the range [0,1.5la]. The rotation angle for each axes is then independently randomly sampled from the interval [−10t, 10t] degrees. The scale factor for each axes is in- dependently randomly sampled from [1 − 0.1t, 1 + 0.1t]. To obtain a translation vector we randomly sample from the surface of a sphere with its radius value randomly sampled from [0, smax t], where smax is the largest dimension of an input network, as before. After a global stroke deformation, we apply a random translation for each stroke vertex. The translation vector is randomly sampled from the disk with a radius rvi , which lies in the plane orthogonal to the stroke direction in the vertex vi. The radius rvi is randomly sampled from the range [0, 0.1lalstroke], where lstroke ∈ [0, 1] is the length of the current stroke. We extend both ends of each stroke by p which value is randomly sampled from [0, 0.1smax ], to reproduce human strokes appearance.

After global and local stroke deformations, we apply 3D spline interpolation, which results in smoother strokes, matching the appearance of human strokes.

For our experiments, we generate 5 synthetic datasets with 5 levels of abstraction, [0.0,0.25,0.5,0.75,1.0], and obtain two mixed datasets by mixing sketches of all abstraction levels or only the sketches with 3 abstraction levels in the middle. shows example sketches obtained with different settings of la parameter.

4. Evaluations

We adopt the following common evaluation measures to evaluate the retrieval performance: Mean Average Precision (mAP), Normalized Discounted Cumulative Gain (NDCG), Nearest Neighbor (NN), which evaluate the ability of a model to discriminate between shape classes, and Top-k ac- curacy, which measure how many of the retrieval tasks have a ground-truth within top k retrieved results.

4.1. Sampling and rendering strategies

We experiment with two rendering styles for NGVNN and two sampling strategies for PointNet++ .

3D sketch rendering for multi-view network. We generate 12 224×224 orthographic views of each 3D shape and 3D sketch by placing 12 virtual cameras around it every 30 degrees, as was proposed by Lee et al.. The cameras are elevated 30 degrees from the ground plane. For both 3D shapes and 3D sketches we experiment with two types of rendering styles: Phong Shading and depth maps. For 3D sketches we represent each line as a 3D tube.

Point cloud sampling. To get the point cloud representation, we first sample 10000 points from shapes and 3D sketches. For shapes we use Monte-Carlo sampling4 and for sketches we use equidistant sampling. We then adopt two types of sampling from the initial 10000 points to obtain a sparse set of 1024 points: random sampling or uniform sampling. The uniform sampling is obtained with the farthest point sampling. The sparse sets are obtained on-the-fly and thus might differ from one iteration to another.

4.2. Effect of sampling and rendering.

We compare different sampling methods for PointNet++ and rendering styles for NGVNN, when training with the classification loss and the triplet loss on the pairs of 3D shapes and 3D sketches rendered with la set to 0.5. the uniform sampling for PointNet++ and depth rendering for NGVNN perform best, so we use these settings for the rest of the experiments.

It can be seen that point-based representation (dashed lines) by far outperforms the multi-view representation (solid lines) when we use the Siamese architecture for NGVNN. We thus for the rest of experiments use the heterogeneous architecture for NGVNN.

5. Conclusion

In this work we proposed the problem of 3D VR-sketch to 3D model retrieval. We first introduced a purpose-built VR environment to collect VR-sketches and conduct retrieval. We then collected a set of 3D human sketches for a subset of two shape classes from ModelNet10. We further proposed an approach for generating synthetic 3D sketches with different levels of abstraction, and demonstrated that the methods trained on our synthetic data generalize well to human sketches. Via a series of comprehensive evaluations, we find that compared to 3D shape-based retrieval, point-based shape representation is advantageous over the multi-view representation. At last, we propose an architecture with an additional sketch regularizing branch that leads to a superior performance over all the considered baselines, demonstrating the benefit of directly tackling the abstract nature of VR-sketches. We hope to have offered a valid first stab at this new problem.