Multiview 3D Reconstruction

Abhishek Aditya BS
Department of Computer Science
PES University
Bangalore

PES1UG19CS019
abhishek.aditya10@gmail.com

T Vijay Prashant
Department of Computer Science
PES University
Bangalore

PES1UG19CS536
tvijayprashant@gmail.com

Vishal R
Department of Computer Science
PES University
Bangalore

PES1UG19CS571
vishalramesh01@gmail.com

Yashas KS
Department of Computer Science
PES University
Bangalore

PES1UG19CS589
yashas120@gmail.com

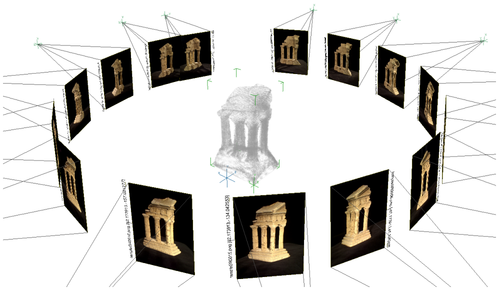
I. Introduction

Creating a 3D model from 2D images that is as realistic as possible is one of the fundamental issues of image-based modeling and computer vision. 3D reconstruction from multiple images is the creation of three-dimensional models from a set of images.The goal of multiview 3D reconstruction is to infer geometrical structure of a scene captured by a collection of images. Usually the camera position and internal parameters are assumed to be known or they can be estimated from the set of images. Recently many applications have emerged, such as autonomous driving and augmented reality, which rely heavily upon accurate 3D reconstructions of the surrounding environment. These reconstructions are often estimated by fusing depth measurements from special sensors, such as structured light, time of flight, or LIDAR, into 3D models. While these sensors can be extremely effective, they require special hardware making them more cumbersome and expensive than systems that rely solely on RGB cameras. Furthermore, they often suffer from noise and missing measurements due to low albedo and glossy surfaces as well as occlusion. By using multiple images, 3D information can be (partially) recovered by solving a pixel-wise correspondence problem.

II. Images from Dataset

Input: Multi-view Images
Output: 3D Reconstruction
ground truth

ShapeNet dataset



ETH Zurich custom dataset

III. Literature Survey

Reconstruction of texture-less objects with unknown surface reflectance is a challenging task in multiview 3D reconstruction. The most challenging task being establishing cross view correspondences where photometric constancy is violated. Ziang Cheng et al., in their paper [1] proposed that the problem can be well posed by multiview geometrical and photometrical constraints and can be solved using a small number of viewpoints. The problem can be formulated as a joint energy minimization over the object's surface geometry and surface reflectance. This energy is highly non-convex in nature and an optimization algorithm was designed that is able to reconstruct the global shape of the object as well as the object reflectance.

The overall setup used by the authors in [1] included a camera setup with a fixed source of light. This point source of light is attached to the camera separated by a small distance from the camera's center. The nature of an object's surface is assumed to be smooth which allows normal vectors to be defined almost anywhere on the object's surface. To recover from unknown bidirectional reflectance distribution function (BRDF) and 3D shape from multiple views of an object, authors use a parameterized BRDF function. The energy function can be formulated as a weighted sum of photometric, shape and BRDF energies. Parametrized BRDF allowed the authors to approximate many real world BRDF functions. These wide range of BRDF functions can be represented as a linear combination of compact BRDF bases with high accuracies in log space. Experimental results showed good object reconstruction that preserves optimal shape of the object as well as its reflectance.

Using an unknown camera to create a 3D volumetric reconstruction from two or more views of a scene raises numerous challenges. While this topic appears to be simple for humans, it presents several hurdles for computers since it needs concurrently rebuilding things across multiple viewpoints while also determining their connection. Unfortunately, present algorithms are not sufficient to the task of volumetric reconstruction from two unknown camera views: this approach necessitates reconstruction as well as pose estimation. Qian et al. in their paper [2] proposed a novel method for estimating reconstructions, as well as distributions over camera/object and camera/camera transformations and an inter-view object affinity matrix. This data is then combined and reasoned through to come up with the most plausible explanation for the scene.

The approach in their paper [2] consists of 3 steps. The two RGB picture inputs are sent through two branches that extract evidence, which is then fused together to get a final conclusion. Object branch, the first network, is a detection network that generates a collection of objects in terms of voxels as well as a transformation into the scene. An object embedding is predicted which will be utilized to construct an affinity matrix between items across pictures. The second network, camera branch, is a siamese network that forecasts a distribution between the camera's translations and rotations. Lastly, the stitching stage analyses the network evidence and generates a final forecast. However, if the image pair is unclear or there are too many comparable things in the scene, the method fails. The random search across object correspondences also limits the stitching step. All conceivable correspondences cannot be explored due to the factorial increase of the search space.

Early works for 3D reconstruction mostly use feature matching between different views of an object. However, the performance of such methods largely depends on accurate and consistent margins between different views of objects and are thus vulnerable to rapid changes between views. Additionally, these methods are not suitable for single-view 3D reconstruction, where only one view of an object is available. The advances of deep learning have shed some light on neural network-based approaches for 3D reconstruction. On the one hand, some researchers formulate 3D reconstruction as a sequence learning problem and use recurrent neural networks to solve the problem. On the other hand, other researchers employ the encoder-decoder architecture for 3D reconstruction. In the proposed method [3] the model is capable of performing end-to-end single and multi-view **3D RE**construction with **TR**ansformers. **3D-RETR** uses a pre trained Transformer to extract visual features from 2D images. 3D-RETR then obtains the 3D voxel features by using another Transformer Decoder.

3D-RETR model from [3] consists of three main components a Transformer Encoder, a Transformer Decoder, and a CNN Decoder. The Transformer Encoder takes as input the images, which are subsequently encoded into fixed-size image feature vectors. Then, the Transformer Decoder obtains voxel features by cross-attending to the image features. Finally, the CNN Decoder decodes 3D object representations from the voxel features.The Vision Transformer uses image xi as input and splits the image into B2 patches. The corresponding patch is embedded at each time step by first transforming it into a fixed-size vector, which is then added to positional embeddings.In the Decoder, the M3 learned positional embeddings are used as inputs and the Transformer Encoder outputs are cross-attended by the Transformer Decoder. This decoder decodes all input vectors in parallel, rather than auto regressively. The Transformer Decoder produces voxel features that are fed into the CNN Decoder, which generates voxel features. The model uses the ShapeNet and Pix3d datasets to evaluate the model using the Intersection of Union (IoU) provides satisfactory results. The 3D-RETR model used in [3] is more efficient than the previous models, as 3D-RETR reaches better performance with much fewer parameters. The 3D-RETR can be further improved by using additional transformers such as Performer, Reformer, etc.

Yariv et al. [4] introduced the Implicit Differentiable Renderer (IDR), an end-to-end neural system that can learn 3D geometry, appearance, and cameras from masked 2D images and noisy camera initializations. Considering only rough camera estimates allows for robust 3D reconstruction in realistic scenarios in which exact camera information is not available.The main advantage of implicit neural representations is their flexibility in representing surfaces with arbitrary shapes and topologies, as well as being mesh-free. Proposed architecture IDR represents the color of a pixel as a differentiable function in the three unknowns of a scene: the geometry, its appearance, and the cameras. Here, appearance means collectively all the factors that define the surface light field, excluding the geometry, i.e., the surface bidirectional reflectance distribution function (BRDF) and the scene's lighting conditions.The model learns 3D geometry of the world from the abundant data of 2D images. In particularly high quality of 3D reconstruction of objects and scenes using only standard images. e. One limitation of our method is that it requires a reasonable camera initialization and cannot work with, say random camera initialization.

IV. Progress

Currently the literature survey and understanding of the problem of Multiview 3D reconstruction is complete. Although implementation has started it will take the remainder of the time till Task 4 to present an implementation. The goal for the next submission is a initial implementation of the model.

References

1. Cheng, Ziang, Hongdong Li, Yuta Asano, Yinqiang Zheng, and Imari Sato. "Multi-view 3D Reconstruction of a Texture-less Smooth Surface of Unknown Generic Reflectance." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16226-16235. 2021.

2. Qian, Shengyi, Linyi Jin, and David F. Fouhey. "Associative3d: Volumetric reconstruction from sparse views." In European Conference on Computer Vision, pp. 140-157. Springer, Cham, 2020.

3. Shi, Zai, Zhao Meng, Yiran Xing, Yunpu Ma, and Roger Wattenhofer. "3D-RETR: End-to-End Single and Multi-View 3D Reconstruction with Transformers." *arXiv preprint arXiv:2110.08861* (2021).

4. Yariv, Lior, et al. "Multiview neural surface reconstruction by disentangling geometry and appearance." *Advances in Neural Information Processing Systems* 33 (2020): 2492-2502.