3D Reconstruction from 2D Image Using Neural Network

Sun Yul Lee

Summary

- Description of Problem

With the rising of VR and AR technologies, people are trying to model objects in the real world in virtual space. With the success of the neural network in extracting features of objects in 2D image, I am trying to address the possibility of using features from 2D image to render 3D models. If the features can represent 3D features of objects, 3D modeling process would be started from taking picture of objects in real world instead of measuring properties of the objects. Further, it would be helpful to rendering complex object with the image.

- Importance of Problem

In recent years, neural networks have contributed a lot to image processing. Through neural networks, people could allow computers to extract important features from images. By using these features, the computer can classify the objects in the image. The success of the neural networks in image processing brought the possibility of utilizing the neural network in more diverse topics. Recently, various industries need 3D modeling. Technologies such as VR or AR requires modeling objects in virtual space. Further, entertainment industries such as movies and games require 3D modeling. It takes a lot of effort and time to design objects in the real world in a computer. If a 3D model can be created using only an image of the object in the real world with a neural network, the neural network model could have a significant impact on the industries.

- Your proposal

  I am trying to build a neural network model which can produce 3D modeling of an object, or a space from 2D image. I will research and study methods such as depth estimation, layout estimation, and semantic segmentation which can be helpful for extracting important features from 2D image. Also, I will research methods which can utilize features from 2D image to represent 3D features of the objects. Eventually, I will try to construct a deep neural network model which can generate 3D modeling of the objects in the image.

- Originality

  There are many papers on using neural network to reconstruct 3D model of object in 2D image. Also, there are many related topics that researchers try to utilize neural network such as layout estimation, pose estimation, depth estimation, and so on. However, during the search, I found that few studies tried to combine related topics to solve the problem. In addition, most research focused on modeling a single object in the image. In the project, I am planning to combine layout estimation, and depth estimation to generate 3D modeling of objects or scene in the 2D image.

List of Goals

- Study and implement CNN model to classify objects in image
- Study and implement depth estimation in image
- Study and implement semantic image segmentation
- Study and implement layout estimation in image
- Build a neural network model for converting 2D image into 3D modeling.

Literature Review

The area of converting 2D image to 3D modeling has been an interesting topic. Yotam Gingold at el[1] tried to provide a system for creating 3D modeling from 2D sketches. They divided a 2D sketch into several cylinder-shaped objects and generate 3D modeling by connecting them into one piece. The method was quite successful in providing a guideline for converting 2D sketches into 3D modeling. But, since the method targeted free-form surfaces, it could not model surfaces with edges or flat surfaces. In addition, people still need modeling process using tools to create 3D modeling.

In recent years, neural networks have brought many advances in image processing. Most researchers are trying to implement neural networks to learn features from 2D image to produce outcome which they want to achieve. Not only the area of 3D reconstruction from 2D image, but also the area of image processing related to the topic achieved significant outcomes from implementing the neural networks. Layout Estimation is a part of 3D reconstruction. Researchers in the topic try to reconstruct 3D structure of a room from a 2D image. Most recent publications in the area trained a neural network to predict edges and room surfaces from a 2D image. One research in the layout estimation approached the problem as regression and tried to label each edge in the image.[2] With labeled edges, the model extracts topology anchor points and constructs the 3D layouts of the scene in 2D image. The recent research on layout estimation with the neural network uses 11 types of room layout. RoomNet[4] is a neural network model which predicts key points in the room layout and a type of the room in the 2D image. Another research separated layout estimation process into two processes: edge mapping and segmentation mapping.[5] The neural network model in the work makes two predictions and combines them to generate layout estimation result. The work shows that using two estimates can enhance the training of the neural network. Those papers show a significant success of utilizing neural network in layout estimation. However, there are limitations in the layout estimation.[2] With the 2D image, the model does not know the distance from the camera center to the layouts of a room which can increase inaccuracy of the model. Also, defined types of the room are not contained rooms with non-orthogonal walls. Unlike forementioned papers which tried to estimate 3D layout of room, one recent research tried to generate top-view layout of road scene from a single 2D image.[17] It utilized neural network to project the features of frontal view to top view. Also, the neural network learns the correlation between the features of frontal view and top view. With learned features, the model produces the top-view layout of road scene. The paper shows a possibility of improving 3D reconstruction by proposing a method to generate one viewpoint image from another viewpoint image.

Depth estimation is another part of 3D reconstruction. Researchers in the topic try to predict the depth value of each pixel with a given 2D image. Depth estimation also started to utilize neural network recently after the success of neural network in image processing. One research tried to merge depth estimation with semantic segmentation.[20] The paper mentioned that the depth estimation and semantic segmentation are related to the property of perspective geometry. The model produces semantic segmentation in the input 2D image. Then it tries to align depth data to the segmented object to predict the depth of pixels in the image. Connecting two fields, the model shows a significant performance. However, there is a drawback since the depth estimation is relied on alignment and semantic segmentation get effect from resolution of the image. Another research tried to approach the problem with unsupervised model.[19] Unlike the previous paper, two image is given to the model in training. The model tried to learn depth feature from the left and right images captured at the same moment in the time. The research showed that depth feature can be captured from two image with different viewpoints. Also, it provided a method to reduce amount of data for modeling the neural network. One research tried to improve the performance of neural network by applying different strategy in training.[18] The model utilized multiple dilated convolution layers to extract multi-scale information. With the extracted information, the model produces the output images. The paper is not directly related to the geometric topic. However, it shows that the performance of the neural network can be improved by many factors. Although the depth estimation achieved high improvement with the neural network, there are still some obstacles such as predicting information for the occluded parts in the 2D image.

After the success of neural network in image processing, a lot of research in the area of 3D reconstruction from 2D image started building various neural network models. Although there are many significant neural network models for 3D reconstruction, the main frames of the models are similar. Most models target to reconstruct a single object in a 2D image to a 3D model. In the training phase, models learn 3D features of each object class and utilize the features in reconstruction. However, each model has different components and approaches to solve the problem. Im2Struct tried to automatically recover a cuboid structure of the object.[8] In the training, the model trains the structure masking network to learn shape feature of objects. Then the model learns part relations including connectivity and symmetry. With the learned features, the model generates 3D model of the object. In another paper, Pix3D is proposed to provide dataset and methods for 3D reconstruction.[9] Researchers contain labeled key points on both 2D images and 3D shapes so that they can be aligned in reconstruction. The paper addressed that the point alignment is one of key parts in the 3D reconstruction. Also, it emphasized that the data structure can give effect on the performance of neural network models. 3D-PRNN is proposed to synthesize multiple plausible shapes composed of a set of primitives.[6] The model tries to represent 3D shape as a collection of simple parts. It generates primitive set of the shape in the image and combine them to create 3D model. The model can generate 3D model. However, since the model generate primitives, the model cannot generate shape in circle and fails to generate accurate 3D model of complex shapes. Some papers tried to build model which can take both single-view image and multi-view images as input. 3D-R2N2 is proposed to learn a mapping from images of objects to their underlying 3D shapes from a large collection of synthetic data.[15] The model reconstructs an object in the form of a 3D occupancy grid with one or multiple images of the object. It uses recurrent neural networks to fuse multiple feature maps extracted from input images. But using recurrent neural networks can lead model to fail in fully exploit input images to reconstruction results. To solve the problem, Pix2Vox was proposed. Pix2Vox also takes one or more images to reconstruct 3D model.[16] The model does not use recurrent neural network. Instead, the model uses context-aware fusion module to select high-quality reconstructions for each part of the object. Unlike other research which tried to build 3D model from 2D images, PointOutNet tried to generate point cloud coordinates from the image.[21] The researchers saw advantages of point cloud in geometric transformations and deformations. They addressed a new concept of building a network for point set prediction. The result of the paper shows that point cloud can be generated from a 2D image, and it can well represent the 3D shape of the object.

Some neural network models tried to represent geometric factors. One paper tried to recover the 3D structure in terms of a layout and set of objects in terms of shape and pose.[3] The model first predicts the layout in the image to represent the enclosing surfaces of the scene. Then it detects objects in the image and predicts 3D model of the objects. The shape of predicted objects is represented as voxel occupancy grid. In another research, 3D-RCNN is proposed to recover 3D shapes and poses of all object instances within a given image.[7] The model predicts bounding box, pose, object centric projection and shape parameters to predict the 3D shape of objects. The researchers utilized differentiable render-and-compare loss to learn 3D shape and pose from 2D images.

There are some papers which tried to perform 3D reconstruction with 2.5D sketches. Researchers brought an idea that reconstructing 3D model is easier with 2.5D sketch than with 2D image. With the idea, MarrNet was proposed.[11] The model has three processes. First it performs 2.5D sketch estimator to predict depth, surface normal, and silhouette images of the object. With the 2.5D images, it predicts 3D object shape through 3D estimator. Then it aligns the estimated 3D shape and the estimated 2.5D sketches. Another paper also combined 2.5D estimator and 3D estimator and built ShapeHD.[13] The main frame of the model is the same as MarrNet. The difference is that ShapeHD introduced a method of regularizing 3D shape if the shape is unrealistic. When the model detects unrealistic shape, priors are used to penalize the model.

One research tried to optimize the 3D reconstruction model with viewpoint labeling.[14] The model is semi-supervised since the viewpoint labeling uses training data but there is no test data to check correctness. The model tried to label the viewpoint of the input image by comparing it with the images with a known viewpoint. The result of the paper shows that with the label of viewpoint, the model can achieve the better performance. Viewpoint labels help the model to align the 3D shape with the data. This addresses the importance of alignment in the 3D reconstruction. Another research tried to improve 3D reconstruction via prior optimization.[12] Most 3D reconstruction models generate priors of shapes in the training phase. However, fixed priors could have poor performance on unseen data. In the paper, researchers tried to optimize the learned prior and latent code according to the input physical measurements after the training. They also showed several applications which can utilize the optimization. Since the 3D reconstruction topic is connected to the neural network topic, it shows that improving the neural network model is also important to improving the performance of 3D reconstruction.

Some papers suggest possible area which the 3D reconstruction can be utilized. Image-based virtual try-on (VTON) requires 3D reconstruction from 2D image of clothes to transform the image to fit on the human images with different poses.[10] the researchers first align clothes image on the silhouette of standard person model. Then they reconstruct 3D model of person with the clothes on and transform the clothes model corresponding to the pose of the person model. Another field that the 3D reconstruction can be implemented is autonomous driving. One paper tried to perform 3D object detection using 3D reconstruction.[22] They first trained model to get position and depth information and generate 3D point cloud. Then they estimated 3D box of the objects in the image to extract 3D data in each box. As the papers show, 3D reconstruction from 2D image can be merged into other field to provide solution to each other field.

Merging with t he neural network, 3D reconstruction has made rapid progress in recent years. Many studies are trying to solve the problem using neural networks. Most neural network models have a encoder-decoder structure to align 2D features to 3D features. But each model proposes different data representation, learning methods, optimization and so on and is continuously developing. Although the 3D reconstruction achieved significant success with neural network, there are still some problems such as representing occluded parts of objects in 2D image that needs to be solved.

References

[1] Y. Gingold, T. Igarashi, and D. Zorin. Structured annotations for 2D-to-3D modeling. In ACM SIGGRAPH Asia, 2009.

[2] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang and F. Xu. 3D Room Layout Estimation from a Single RGB Image. In IEEE Transactions on Multimedia, vol. 22, no. 11, pp. 3014-3024, Nov 2020

[3] S. Tulsiani, S. Gupta, D. Fouhey, A. Efros, and J. Malik. Factoring Shape, Pose, and Layout From the 2D Image of a 3D Scene. In CVPR, 2018

[4] C. Lee, V. Badrinarayanan, T. Malisiewicz, and A. Rabinovich. RoomNet: End-To-End Room Layout Estimation. In ICCV, 2017

[5] W. Zhang, W. Zhang and J. Gu. Edge-Semantic Learning Strategy for Layout Estimation in Indoor Environment. In IEEE Transactions on Cybernetics, vol. 50, no. 6, pp. 2730-2739, June 2020

[6] C. Zou, E. Yumer, J. Yang, D. Ceylan, and D. Hoiem. 3D-PRNN: Generating Shape Primitives with Recurrent Neural Networks. In ICCV, 2017

[7] A. Kundu, Y. Li, and J. Rehg. 3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare. In CVPR, 2018

[8] C. Niu, J. Li, and K. Xu. Im2Struct: Recovering 3D Shape Structure from a Single RGB Image. In CVPR, 2018

[9] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. Tenenbaum, and W. Freeman. Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In CVPR, 2018

[10] M. Minar, T. Tuan, H. Ahn, P. Rosin, and Y. Lai. 3D Reconstruction of Clothes Using a Human Body Model and Its Application to Image-Based Virtual Try-On. In CVPR, 2020

[11] J. Wu, Y. Wang, T. Xue, X. Sun, W. Freeman, and J. Tenenbaum. Marrnet: 3D Shape Reconstruction Via 2.5D Sketches. In NIPS, 2017

[12] M. Yang, Y. Wen, W. Chen, Y. Chen, and K. Jia. Deep Optimized Priors for 3D Shape Modeling and Reconstruction. In CVPR, 2021

[13] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. Freeman, and J. Tenenbaum. Learning Shape Priors for Single-View 3D Completion and Reconstruction. In ECCV, 2018

[14] I. Laradji, P. Rodriguez, D. Vazquez, and D. Nowrouzezahrai. SSR: Semi-supervised Soft Rasterizer for single-view 2D to 3D Reconstruction. In ICCV, 2021

[15] C. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In ECCV, 2016.

[16] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang. Pix2Vox: Context-Aware 3D Reconstruction from Single and Multi-View Images. In ICCV, 2019

[17] W. Yang, Q. Li, W. Liu, Y. Yu, Y. Ma, S. He, and J. Pan. Projecting Your View Attentively: Monocular Road Scene Layout Estimation via Cross-View Transformation. In CVPR, 2021.

[18] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In CVPR, 2018.

[19] C. Godard, O. Aodha, and G. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In CVPR, 2017.

[20] L. Ladicky, J. Shi, and M. Pollefeys. Pulling Things out of Perspective. In CVPR, 2014.

[21] H. Fan, H. Su, and L. Guibas. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In CVPR, 2017.

[22] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan. Accurate Monocular 3D Object Detection via Color-Embedded 3D Reconstruction for Autonomous Driving. In ICCV, 2019.