Reconstruction of 3D Objects using Neural Network

Aishwarya G[1], Likhita B[2], Nidhi J[3], Dr M S Patel[4]

[1,2,3]UG student, [4]Professor, Dept. of Information Science and Engineering, Sapthagiri College of Engineering, Bengaluru, Karnataka, India

---

Abstract: - *Motivated from the recent developments of mechanisms that make use of shapes to achieve 3D reconstructions, we present a system built on neural networks. Our network acquires skill to map target images to their respective 3D reconstructions with the help of artificial data. Multi-view images of target is fed to this network as input and after completion of the process the network returns target's reconstructed view which is the form of 3D.This network does not need any notations on the target image for training or testing which makes it different from other recent works. This network is an add-on to the typical LSTM networks that can easily fit in the images of target in a proper way.*

Keywords: **RNN (recurrent neural network), multi-view, reconstruction, 2D Convolutional Neural Network, 3D De-convolutional Neural Network and 3D Convolutional LSTM.**

1. INTRODUCTION

In recent years, image based 3D reconstruction has become an interesting topic for research. The emerging automatic prototyping of 3D object have become one of the revolutionary change in many implementations like the visualization, electronic commerce and many more. Generic models of organs are created using multiple 2-D image slices for the medical industries. This trend has been uplifted now by using more accurate and efficient methods like the 3D acquisition method and 3D printing which is a standardized technology. Moreover, large data of 3D object models are also coupled in the trend of discussion such as ShapeNet. Almost all the methods are limited to some restrictions for 3D reconstruction. Some of them however are: i)The number of views required to view an object should be dense, views must comparatively include a compact baseline. These kind of issues usually exist when reconstruction is expected from either a fewer number of views or just a view. ii) The albedos are meant be rich of non-homogeneous textures and the objects with reflectance functions are looked forward to be non-reflective.

Looking at these kind of restrictions we arrive at a multiple basic technical assumptions. One such assumption is that features can be extracted and mapped using multiple views of the object. It has been demonstrated that the separation of viewpoints by large baseline, results in a problematic situation for establishment of traditional feature correspondences due to the variation in its local appearances.

In additional, the feature matching problem becomes a very difficult task due to specular reflections and lack of texture on objects. So as to avoid the issues regarding large baselines there are methods of 3D volumetric reconstruction like space carving and their possible extensions which are quite popular. Yet these methods presume that the objects are precisely segmented from the framework which may not be true in many cases.

A different advantage is that the information of appearance of the target object and its shape is already available. Therefore this acts as an advantage which makes sure that the reconstruction method is less dependent on observing exact features similar across the views. So the system can work with less number of images and lesser assumptions. For the observations to adapt 3D shape models, sophisticated mathematical equations have been formulated.

This paper is in the same way with a key variation where, we make use of deep convolutional neural networks in order to understand the mapping from 2D objects to their respective 3D shapes, from huge training data rather trying to match a 3D structure before the object observations and adapting to it.

In this paper, however, we hold the capability to learn automatically from deep neural networks, in end-to-end fashion, to reconstruct an object with few images with minimal supervision.

2. LITERATURE SURVEY

Christopher B. Choy and his group presented a network that merges both single-view and multi-view 3D reconstruction into a unified structure. Their approach required least overseeing in testing and training the network. For training purpose, they made use of one or more images as input which used to vary. To be more precise the number of images of target object for one training small training batch is fixed and varies among different small training batches. Shapenet dataset was used for training purpose [1].

Jaesik Park and his group presented a system to extract geometrical characters from an image since it becomes the primary step in reconstructions. Performance is dramatically improved with the help of metric losses. They tried to validate the experiments using both indoor and outdoor data. They demonstrated that FCGF (Fully convolutional geometric features) provides higher accuracy than other methods by showing that it is 600 times faster than previous methods [2].

JunYoung Gwak and his group presented a system which uses front masks for 3D reconstructions. As an add on, they also proposed a system to connect the gap between 2D masks and 3d volumes. They also made a study on artificial images in order to examine every part of the model. Compared to other previous works with 3D supervision this system can generate a finest reconstruction from a weak 2D supervision [3].

XinchenYan and his group prepared a process for learning which is a relationship between 2D and 3D portrayals, a network consisting of encoder and decoder. Without using the 3D data for training this neural network has the ability to predict 3D shape from single view. At the time of learning, using back propagation algorithm a volumetric 3D shape is generated through a single view input. Therefore using a single view image 3D shape can be generated immediately at the time of testing [5].

3. PROPOSED SYSTEM

Every 3D-R2N2 model contains 3 parts namely encoder, decoder and recurrence unit. Here LeakyReLU nonlinearity is arranged after every layer of convolution. The encoder firstly feeds all the features into 3D-LSTM by converting a 127_127 RGB image into its lower dimensional features. The encoded hidden states of the 3D model is fed into the decoder, which in-turn converts this to map. LeakyReLU comes after every layer of convolution. In this paper, there are two versions of 3D-R2N2 used: deep residual network at the bottom and shallow network in the top layer.

The network of 3DR2N2 consists of three components namely: 2D Convolutional Neural Network (2D-CNN), 3D De-convolutional Neural Network (3D-DCNN) and 3D Convolutional LSTM (3D-LSTM) which is a novel architecture. Firstly, our networks component 2D-CNN is used for implementation where one or more pictures of the structure taken from various view-points are fed into the 2D-CNN. Now 2D-CNN encodes these images taken as input (consider as x) within their lower dimensional features (t(x)). 3D Convolutional LSTM [1] of the network is fed with this encoded image input units and is used for either selectively updating of its cell states or for retaining the states by shutting the gates of input. Eventually, 3D probabilistic voxel reconstruction is generated by the 3D-DCNN by decoding the hidden states of units of the LSTM.

Recurrent neural network

A brief review of LSTM (Long Short-Term Memory) networks and GRU (Gated Recurrent Units) which is a slight variation of the LSTM [1] is provided in the below section.

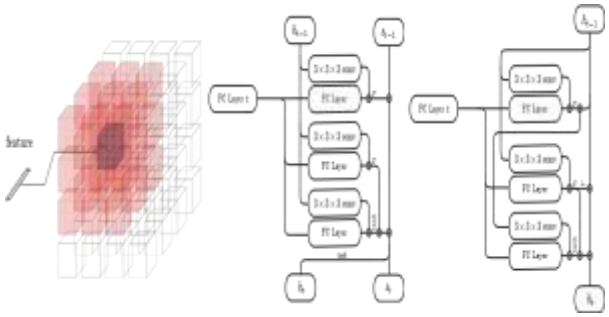
Long Short-Term Memory Unit (LSTM): This unit is one of the highly triumphant implementation of the hidden states of RNN (recurrent neural network). An LSTM unit allows our network to exceed the problem of vanishing gradient by distinctly managing the flow from input to the output. Categorically, there is presence of 4 components in LSTM unit. First component is the memory unit which comprises of a hidden state and memory cell, rest of the components are the gates where each one is used for different purposes. The first gate manages the direction of data flow from the input to the hidden state and this gate is notified as the input gate. The second gate also known as output gate controls the flow from the hidden state to the output. The last gate directs flow from the previously hidden state to the current hidden state and this gate is named as forget gate.

Gated Recurrent Unit: Cho et al has put forward a method which is a slight variant of the LSTM unit called the GRU (Gated Recurrent Unit). An advantage of the GRU over the LSTM unit is that there are lesser computations than that of LSTM. Both the input gate and forget gate are administered by the update gate in GRU. Before the nonlinear transformation a reset gate is appealed by GRU which is one of the other differences between GRU and LSTM.

3.1 Encoder: 2D-CNN

For encoding of the image into its features we make use of a Convolutional Neural Network (CNN). In our paper, Two distinct 2D-CNN encoders are designed which are mentioned in fig 1 below. Fig 1 (a) shows a feed-forward CNN and (b) shows deep residual variation of standard free-forward CNN. Pooling layers, leaky rectified linear units and standard convolution layers are present within the first network which is then lead by a fully-connected layer. Intense residual variation is generated for the initial network inspired by its current studies [1].

According to our work, deep networks optimization procedure speeds up and enhances by addition of residual connections amid convolution layers that are standard. The identity mapping connections are present after each and every two convolutional layer (excluding the 4th pair) of the deep residual variation within the encoder network. We demonstrate the usage of 1*1 convolution in-order to match the various channels following convolutions for residual connections. The completely connected layer is fed with the flattened output of the encoder and this layer squeezes the output fed into a 1024 feature dimension [1].



(a) (b) (c)

Fig 1: (a) input for LSTM unit (b) 3D Convolutional LSTM (c) 3D Convolutional GRU's

3.2 Decoder: 3D De-convolutional Neural Network

An input image segment $x_1, x_2, \cdots, x_T$ is first collected from the earlier process after which the hidden state h_T is processed to the decoder by a 3D-LSTM. This results in rising the resolution of hidden state when applied with 3D unpooling, 3D convolutions and non-linearities [3] until target output resolution is reached. In this network, we are presenting a deep residual version consisting of four residual connections and simple decoder network carrying five convolutions, followed by a final convolution which is similar to that of an encoder. Reformation of final activation $V \in R^{Nvox \times Nvox \times Nvox \times 2}$ to that of occupancy probability $p_{(i,j,k)}$ of the voxel cell at (i,j,k) is carried out when activation outstretches the resolution of the target output utilizing a voxel-wise softmax[1].

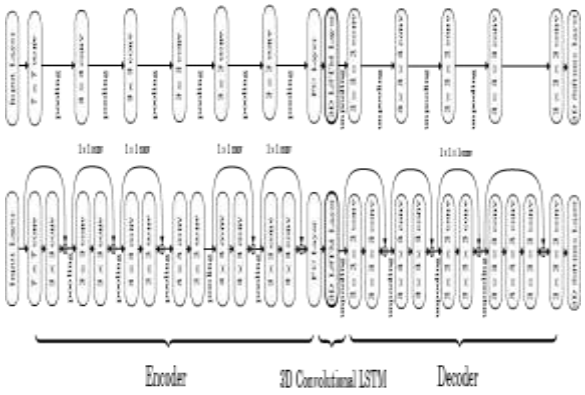


Fig 2: Representation of internal working

3.3 IMPLEMENTATION

Data augmentation: We have used ground truth voxel occupancy maps in training and 3D CAD copies for getting input images. Here, we have first taken the CAD copies with a clear background and then raised the input images with random crops. Also, we slightly colored the copies and arbitrarily translated the images. Also all view-points were used as a sample arbitrarily.

Training: Here, we use changing input lengths that ranges from one image to multiple images. To be more specific, the number of views for each training example was kept constant for each example in training within a single mini-batch, but the length of input training examples varied with different mini-batches arbitrarily. This allowed the reconstruction of both single-view and multi-view by the network. We calculated the loss at the end of an input sequence in order to save both memory and computational power in the training. The other way, by taking the hidden states of the LSTM units at the time of testing, we could retrieve the intermediate reconstructions at each step.

Network: Input image's size was set to 127 × 127. The reconstructed output's size was 32 × 32 × 32. We used 60,000 iterations to train the network where the batch size is 36 except for [Res3D-GRU3], which used a batch size of 24 to fit in an NVIDIA Titan X GPU. The slope of the leak was set to 0.1 for the LeakyReLU layers throughout the network. This method uses Adam for the SGD update rule and Theano to implement our network. We initialized all weights except for Long Short Term Memory weights using MSRA. We used an unitary matrix for initialization in LSTM and SVD to decompose a random matrix.

3.4. DATASETS

ShapeNet: The dataset of ShapeNet is a vast information rich repository which contains 3D CAD models. They are arranged as stated by WordNet hierarchy. This paper makes use of ShapeNet dataset subset that contains 13 major categories and 50,000 models. The dataset is split into 2 parts: testing sets and training sets, where 4/5 of it is used for training and the remaining 1/5 of it is used for testing. So we name these two set of data as ShapeNet training set and ShapeNet testing set all over the process.

Online Products: There are about 23,000 images included in the dataset which are marketed online. Because of ultra-wide baselines, methods like SFM and MVS fail with images. We only make use of datasets for qualitative.

MVS CAD Models: We accumulate different types of higher quality CAD models in order to differentiate our process with other multi view methods.

Each and every CAD models are put on a texture-rich paper and models have texture-rich surfaces to aid the camera localization of the multi view methods.

Metrics: In order to estimate the grade of the recinstruction we used two of the metrics.The first metric is voxel Intersection over union and it was in the middle of 3D voxel reconstruction structure and its observed voxelized model.
More officialy,

$$IoU = \sum_{i,j,k} \left[I(p_{(i,j,k)} > t) I(y_{(i,j,k)}) \right] / \sum_{i,j,k} \left[I\left(I(p_{(i,j,k)} > t) + I(y_{(i,j,k)}) \right) \right] \quad (17)$$

where variables are defined. Here t represents voxelization threshold and $I(\cdot)$ is an indicator function. Greater IoU values represents good reconstructions. Secondary metric is the cross-entropy loss. It reports that lower loss values indicate bigger reconstruction confidence [1].

4. RESULT

Datasets of online products are used for the qualitative testing of our network. Fig 2 shows the reconstruction of the 3D images for our sample objects. Considering only the synthetic data as the training samples we are able to construct real world objects. With increase in number of views of object, the better it is for our network for reconstruction. Now let us consider the dataset of the three seater sofa from the fig 2 as an example for our examination.

The various exploratory results obtained for the reconstruction of 3D images is depicted in the figures below.

Fig 3: Datasets of online products. Input sequence (left to right) are viewed in top rows. Reconstruction of the 3D image at every time step are viewed in bottom rows.

Noticing the side view of the sofa the network constructs the object as a one seater sofa, but after viewing the front view our network reconstructs the object by modifying it into a three seater sofa. The result of reconstruction directly reflects on the number of views and the observations made.

We can also use a single real world image for reconstruction of 3D object for the network. Here we are experimenting the performance of the network in comparison to Kar et al recent works for single view reconstruction. The result of our method of approach exceeds in every categories compared to that of Kar et al.

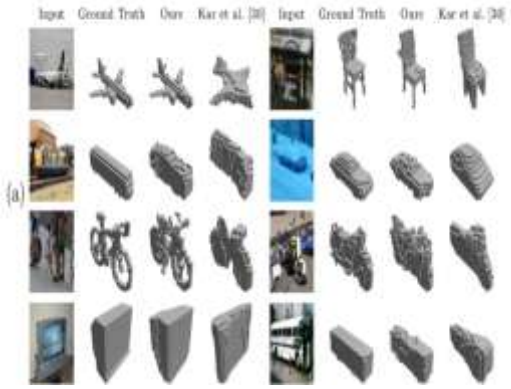


Fig 4: Reconstruction of 3D images for single view using PASCAL VOC dataset

The ShapeNet testing sets are used for the evaluation of our network. There are 8725 models with 13 main categories within the testing set. We select random 5 views

of the object and report the interaction over union (IoU) and loss of the object to the network. The IoUs of all the 13 main categories are depicted for reconstruction of the model in the figure 3. The network first detects the quality of the various views presented for reconstruction model. With increase in number of views the reconstruction quality will also gradually increase as notified in figure 3. But the increasing quality differ for various objects based on the different types of objects. Cars, speakers and cabinets have the highest performance of reconstructing due to their bulky shape and are less shape different when compared to other models. Chairs, lamps have comparatively less reconstruction performance due to their sizes and high shape differences.

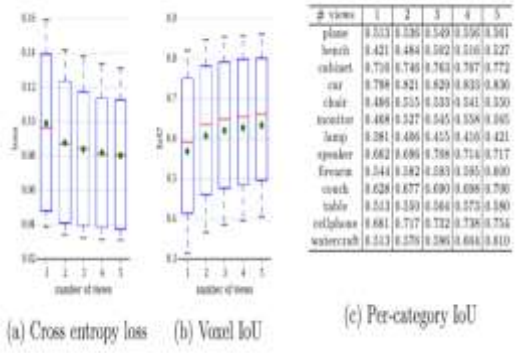


Fig 4: (a) Cross entropy loss (b)Usage of ShapeNet dataset for Multiple view reconstruction of the model. (c)Pre-category reconstruction of ShapeNet dataset using our model.

Our method does not need to train and test each and every categories, as it trains and produces the reconstructed objects without even knowing the object classification. Segmentation of the objects and label of the object doesn't play a key role in reconstruction and is not necessary. The result of reconstruction does not depend upon the description of the object and can be applied even for unknown 2D objects.

5. ADVANTAGES

- Reconstruction of 3D objects using multi view of 2D images for real world images
- Reconstruction of model with minimum number of images instead of complete 360 degree view.
- 3D construction of any model or machine without help of any design engineers.

6. APPLICATIONS

- Face Detection.
- Language Modelling and Generating text.
- Translation of Machine.
- OCR Applications as Image Recognition.
- Prediction problems.
- Text summarization.

7. CONCLUSION

In our work we propose new methodologies for generation of 3D objects from their underlying 2D images. The 3D images are reconstructed without losing the actual dimensions of the object. We demonstrate the usage of both single view and multiple view procedures for reconstruction in our paper. Reconstruction is possible for online product datasets and ShapeNet training sets, which improves the networks performance with increase in number of views. Further it is noticed that quality of the reconstructed object increases with increase in its views. The network performance for the reconstruction varies for different categories of images based on shape, size and variation of the structure. We analyze that single view reconstruction outperforms compared to kar et al model for real time objects. It is possible to reconstruct unknown object without considering its segmentation and key-label.

8. FUTURE WORK

Future work includes enhancement of the network to work on video processing and can be expanded for more applications like entertainment, training and educational system and games in complex environment.

REFERENCES

[1] Christopher B. Choy, Kevin chen, Silvio savaraese, "A unified approach for Single and Multi – view 3D reconstruction", Stanford University, 2019.

[2] Christopher Choy, Jaesik Park and Vladlen Koltuully, "Fully Convolutional geometric fratures", ICCV 2019.

[3] JunYoung Gwak, Christopher B. Choy and Manmohan Chandraker, "Weakly supervised 3D Reconstruction with Adversial constraint" Stanford University, 2019.

[4] Xian-Feng Han, Hamid Laga, Mohammed Bennamoun, "Image based 3D object reconstruction: State-of-the-Art and trends in the Deep Learning Era", in Stanford University, 2019. (pg 9).

[5] Xinchen Yan, Jimei Yang and Yijie Guo, "Perspective Transformer nets: Learning SingleView 3D Object

Reconstruction without 3D Supervision" Stanford University, 2016.
[6] Openmvs: open multi-view stereo reconstruction library (2015), https://github.com/cdcseacave/openMVS, [Online; accessed 14-March-2016]
[7] Cg studio (2016), https://www.cgstud.io/, [Online; accessed 14-March-2016]
[8] A.Dosovitskiy, J.T.Springenberg, T.Brox: Learning to generate chairs with convolutional neural networks. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
[9] Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building rome in a day. In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE (2009)
[10] Anwar, Z., Ferrie, F.: Towards robust voxel-coloring: Handling camera calibration errors and partial emptiness of surface voxels. In: Proceedings of the 18th International Conference on Pattern Recognition - Volume 01. ICPR '06, IEEE Computer Society, Washington, DC, USA (2006), http://dx.doi.org/10.1109/ICPR.2006.1129