Volumetric Spatial Transformer Network for Object Recognition

Min Liu, Yifei Shi, Lintao Zheng, Yueshan Xiong, Kai Xu*  
National University of Defense Technology, HPCL  
*kevin.kai.xu@gmail.com

1. Introduction

Understanding 3D environments is a vital element of modern computer vision research due to paramount relevance in many vision systems, spanning a wide field of application scenarios from self-driving cars to autonomous robots [1]. At the present time, object recognition mainly employs two methods: volumetric CNNs [2] and multi-view CNNs [3] [4]. In this paper, we propose a volumetric spatial transformer network for object recognition. It fills the gap between 3D CNN and 2D CNN for the first time, and provides an end-to-end training fashion. Given a 3D shape, the network can automatically select the best view that maximizes the accuracy of object recognition.

2. Approach

The main idea of our volumetric spatial transformer network is to build up an end-to-end fashion deep neural network. As illustrated in Figure 1, our volumetric spatial transformer network mainly consist of three parts: a 3D CNN, a depth layer and a 2D CNN. For a 3D shape, we first convert the shape into volumetric representation with a (60×60×60) resolution. The 3D volumetric data is then fed into the 3D CNN. In order to mitigate overfitting, we adopt the mlpconv layer from [1]. Our 3D CNN includes 60×60×60 convolutional layers and 60×60×60 fully connected layers. As illustrated in Figure 1, our volumetric spatial transformer network for object recognition. It fills the gap between 3D CNN and 2D CNN for the first time, and provides an end-to-end training fashion. Given a 3D shape, the network can automatically select the best view that maximizes the accuracy of object recognition.

A backpropagation through the depth layer computes loss gradients at input (θ, ϕ) given loss gradients of output (depth image). During backpropagation, each pixel will get a ∂loss/∂d. We know the depth d and (X, Y, Z) can be get from the index map. That means ∂d/∂X, ∂d/∂Y, ∂d/∂Z in Equation 5 can be calculated along with Equation 1 and Equation 4.

Each pixel of depth image will get a ∂d/∂X, ∂d/∂Y, ∂d/∂Z, we just average all the ∂d/∂X, ∂d/∂Y, ∂d/∂Z to get the final loss gradients of (θ, ϕ).

3. Results

To evaluate the performance of view selection, we compare our Spatial Transformer Network against two alternative methods. We use two baseline approaches, i.e., Projected Area which selects the view by maximizing the area of projection of a 3D model, Random which selects viewpoint randomly. The recognition network is trained on 40 object categories, and each category has 250 3D shapes. Figure 3 shows some of the recognition results. It is obvious that our view selection approach by Spatial Transformation Networks outperforms baseline approaches. Figure 4 shows some view selection results.

4. Conclusions

We propose a volumetric spatial transformer network, which is proved to be effective for object recognition. The intermediate result (θ, ϕ) can be used for many other task. We can extend the work by combine recurrent neural network (RNN) with our volumetric spatial transformer network, so it can develop next best view approaches to reduce the recognition uncertainty of 3D objects with a minimal number of views. This work was supported in part by NSFC (61379103, 61572507, 61523003, 61622212).

References