Monocular 3D Reconstruction of Polyhedral Shapes via Neural Network

Mark Beers, Zygmunt Pizlo

Cognitive Sciences, UCI

Many tasks in early human vision have been identified as ill-posed inverse problems. Monocular 3D reconstruction is a particularly difficult inverse problem as there are many 3D scenes that could have generated any given 2D image. In order to select among these 3D interpretations, prior belief must be incorporated that deems one 3D interpretation more likely than another. Shape constancy experiments have illuminated some of the prior beliefs that humans have about the world. In particular, humans accurately reconstruct symmetrical, compact, planar-faced objects, if such an interpretation exists. Prior modelling work has demonstrated that incorporating these prior beliefs into a cost function allows for unique and accurate 3D reconstruction. Unfortunately, cost function minimization is often a slow, iterative process. Yet, humans reconstruct 3D scenes well in a fraction of a second. In this work, we hope to mimic both the accuracy and the speed of 3D reconstruction.

Prior belief is naturally incorporated via regularization as in equation 1. In this equation, X represents a 3D shape, A a projection operator mapping a 3D shape to a 2D image, Y a 2D image, and P a set of prior beliefs about X, and λ a quantity governing the relative importance of data versus prior belief. In our case, this equation is simplified by the use of orthographic projection. Under orthographic projection $(X, Y, Z) \rightarrow (X, Y)$ and so any Z we propose for a vertex will yield a 2D image consistent with the input image. Therefore, under orthographic projection, $||AX - Y||^2 = 0$ for all X. We remove this data driven term from equation 1 and expand on our prior beliefs to arrive at equation 2. In equation 2, S_{dist} is a measure of 3D asymmetry, P is a measure of non-planarity of faces, and C is a measure of compactness, a large value of which implies high volume of a 3D object for a given surface area. Therefore, the cost function we have designed will select a reconstruction that is minimally asymmetrical, is maximally compact and has roughly planar faces. These components of the cost function have been selected to emulate human shape constancy performance.

$$E = ||AX - Y||^2 + \lambda||P(X)||^2 \tag{1}$$

$$L_p = (w_s S_{dist} + 1)(w_p P + 1)(w_{c_1} e^{-w_{c_2} C} + 1) \tag{2}$$

In this work we consider 16 vertex polyhedrons similar to those in figure 1, each represented as a graph with each vertex assigned (x, y) coordinates. For iterative, gradient descent based minimization, we fix the depth value of one of these vertices and search (using BFGS, 10 random starting points) for depth values of the remaining 15 vertices to find a cost function minimizing reconstruction. We also implement a graph convolutional neural network (GCNN) that has been trained to minimize equation 2. The neural network uses a RELU activation function between each layer and is trained with the ADAM optimizer. For a given graph and cost function, we compare GCNN reconstruction to explicit optimization reconstruction. We conclude that, conditional on a test graph being similar to a graph in the training set, our network can reconstruct the 3D shape of a polyhedron very quickly, often with only slight degradations in reconstruction quality relative to more traditional iterative methods. We will explore how GCNN generalization performance degrades as we test graphs further from the training data. Figure 1 shows explicit optimization and neural network based reconstructions of two sample shapes. These input shapes and viewing directions are similar to, but not identical to, shapes and viewing directions shown to the network during training.

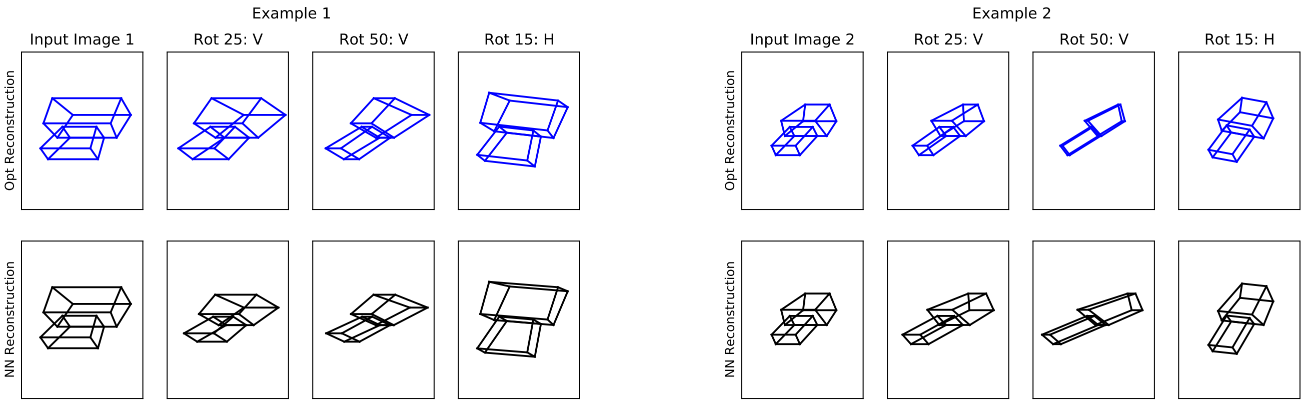

Figure 1: Two examples of 3D shape reconstruction by the NN and by an explicit optimization (Opt). The input image is on the left. The additional 3 views show the reconstructed shape after rotation around vertical and horizontal axes.