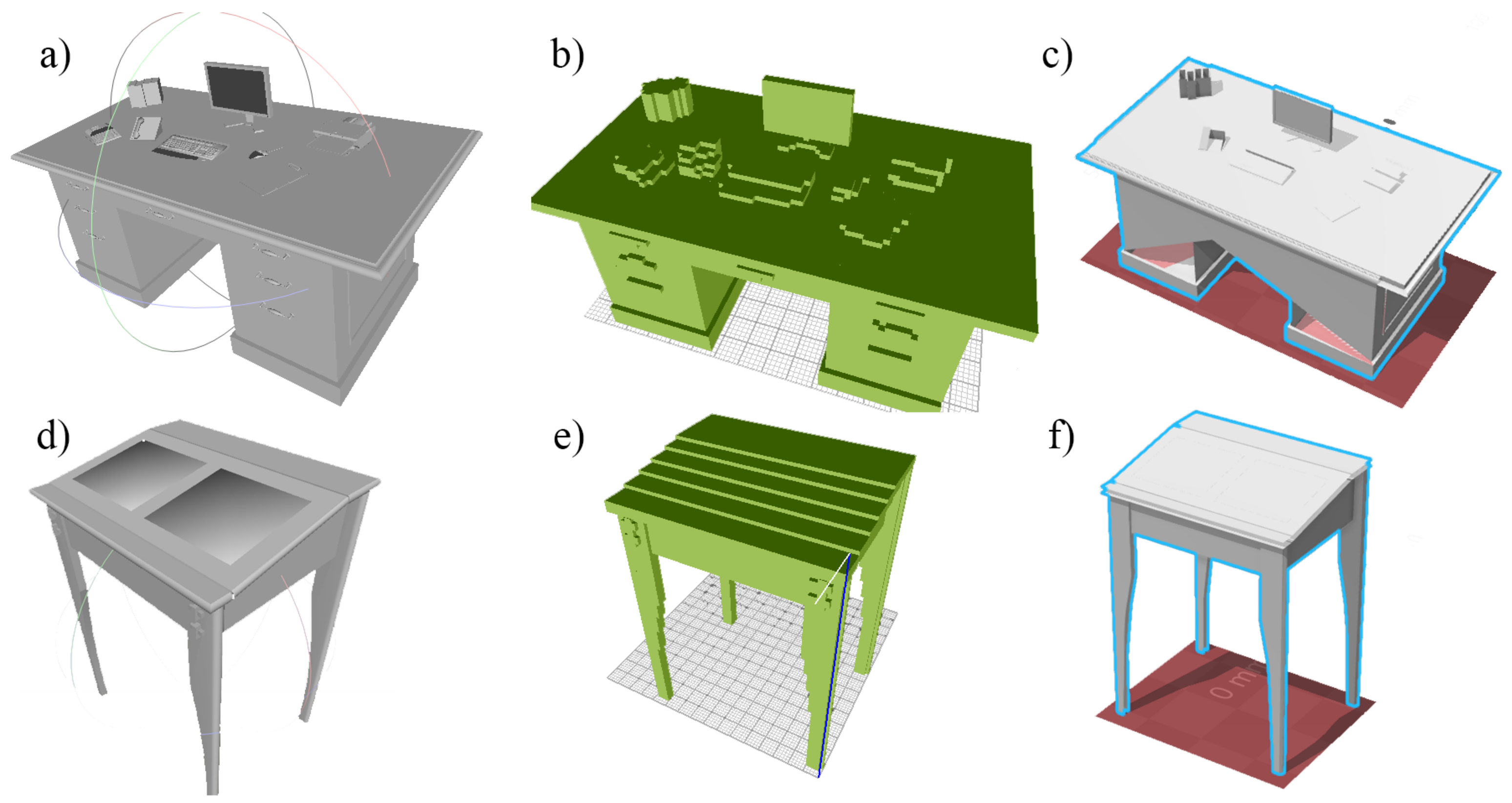
TOWARDS 3D CONVOLUTIONAL NEURAL NETWORKS WITH MESHES

Miguel Dominguez, Felipe Petroski Such, Shagan Sah, Raymond Ptucha
Rochester Institute of Technology, Rochester, NY, USA

Introduction

- Point clouds cannot be classified by traditional CNNs (not an array)
- Voxels have an $O(d^3)$ storage cost
- Treating point clouds as a graph more efficient

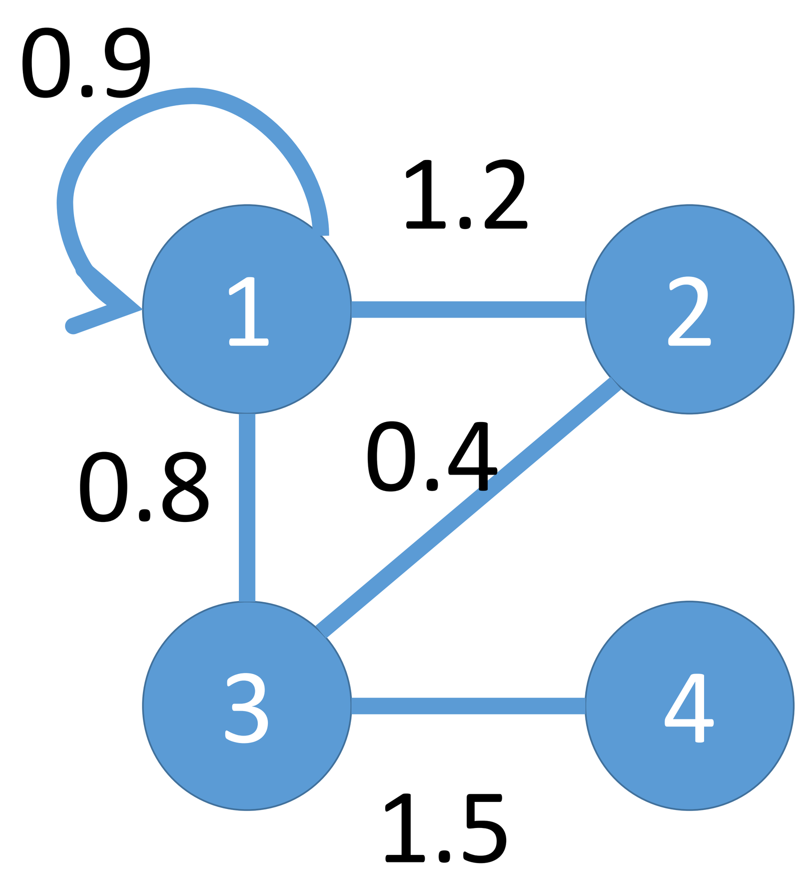
Voxels vs Meshes



- Voxels can be used for accurate classification if heavily downsampled upfront
- Even high-dimension voxels introduce artifacts
- Low-dimensional meshes preserve detail
- Asymptotically more efficient:
  - Naïve implementation: $O(N^2 + N)$
  - Can be $O(N + E)$ with sparse implementation

Proposed Method

- Graphs can be posed as $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{A})$
- $\boldsymbol{V} \in \mathbb{R}^{N\times f}$ are vertices, $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is adjacency matrix
- Can produce filters as polynomials of $\boldsymbol{A}$
- $\boldsymbol{H} = h_0\boldsymbol{I} + h_1\boldsymbol{A} + h_2\boldsymbol{A}^2 \ldots h_k\boldsymbol{A}^k$
- In CNNs, can approximate as a cascade of small filters:
- $\boldsymbol{H_{base}} \approx h_0\boldsymbol{I} + h_1\boldsymbol{A}$
- Filtering $\boldsymbol{V}$: $\boldsymbol{V_{out}} = \boldsymbol{HV_{in}}$
- Stacks of L adjacency matrices can encode multiple features and learn more parameters



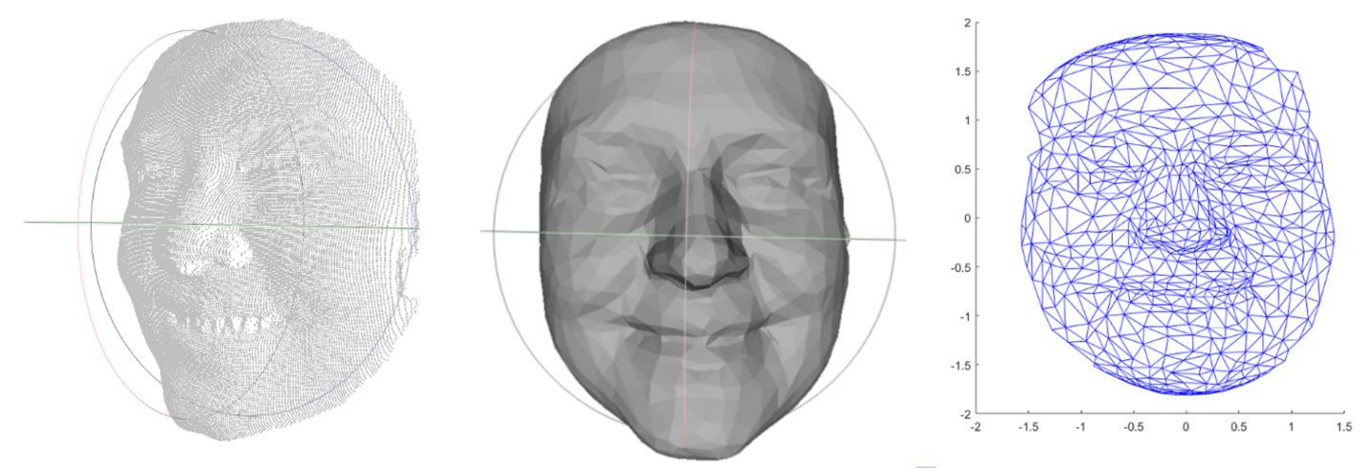
$$\boldsymbol{A} = \begin{bmatrix} 0.9 & 1.2 & 0.8 & 0 \\ 1.2 & 0 & 0.4 & 0 \\ 0.8 & 0.4 & 0 & 1.5 \\ 0 & 0 & 1.5 & 0 \end{bmatrix}$$

$$\boldsymbol{V} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

- Row i of $\boldsymbol{A}$ is an indicator function
- Nonzero values indicate connections to vertex i
- $h_1 0.9(1) + h_1 1.2(2) + h_1 0.8(3)$ is the weighted sum of vertex i "one-hop" neighborhood
- $\boldsymbol{A}^2$ is "two-hop" neighborhood, encodes connections two hops away from vertex i

Results

Task 1: Facial expression recognition on point clouds



Architecture	Accuracy	# Parameters
3x GraphConv16	54.8%	8016
4x GraphConv16	70.0%	10336
5x GraphConv16	67.9%	12656
CNN + Images	**82.2%**	8032

- 5-layer Image CNN may have advantage due to richer input features

Task 2: Modelnet10 3D meshes

Architecture	Accuracy
VRN Ensemble	**97.14%**
ORION	93.8%
4x GraphConv24(ours)	74.3%
Ravanbakhsh,et al. Graph Method*	58.0%

*Only has results on ModelNet40, not Modelnet10

- Limited receptive field, pooling and stride to be added in the future to increase the field
- Contact Miguel Dominguez at mad6384@rit.edu