Viewpoint Invariant Person Classification in RGB-D Data
Alisha Rege (amr6114@stanford.edu)

Purpose
Artificial Intelligence can play a key role in healthcare; however, due to patient confidentiality (HIPAA), we are unable to process this information without putting up some boundaries. This boundary comes in the form of RGB-D data; it prevents us from seeing a face or a distinguishing personal characteristic in videos. This project attempts to detect a person from any viewpoint in Stanford Health’s RGB-D data. The goal is to create a detection system that will be able to identify a person from any viewpoint. This will allow nurses and doctors to sense problems such as if a person suddenly fell or if the person has not moved in days. A 6-layer CNN classifier is used to classify the object.

Dataset
- Dataset from Stanford’s Lucile Packard Children’s Hospital
- RGB-D data from 3 different viewpoints
- Hand-labelled Data
- Previous project created bounding boxes for objects
- Large variations in viewpoints, object appearance and pose, object scale

Dataset Preprocessing
- Translating Video into frames
- Cropping annotations to feed into network
- Resizing all images to 56x56x1
- Batch Size: 50
- Cleaning data

Baseline
- SVM w/ HOG Descriptors
- HOG Descriptors: Slides through image and calculates the number of gradients in certain direction
- Uses NMS – only one major object in certain pixel range
- Skimage Implementation $\min_{\frac{1}{2}} \frac{1}{2} ||w||^2 + \lambda ||w||^2$

Convolutionsal Neural Network Implementation
6 – Layer CNN w/ Dropout:
- Input (56 x 56 x1)
- Convolutional Layer (5x5 convolution w/ 32 Bias) & RELU
- Pooling Layer (2 down samples)
- Convolutional Layer (5x5 convolution w/ 64 Bias) & RELU
- Pooling Layer (2 down samples)
- Fully-Connected Layer (14x14x64 Inputs -> 1024 Outputs)
- Out (1024 Inputs -> 2 classes)

Convolutional Layer:
$x_{ij}^l = \sum_{m=1}^{m-1} \sum_{a=0}^{m-1} \omega_{ab} y_{(i+a)(j+b)}^{l-1} \text{nonlinearity} y_{ij}^l = \sigma(x_{ij}^l)$
- $m_x m_y$ convolutional size
- $x_{ij}^l$ is every pixel selected

Mean Average Precision (mAP)
$mAP = \frac{1}{\text{numLactories}} \sum_{c=1}^{\text{numLactories}} \text{AP}(c)$

Effect of Training Size on Output:

Discussion

Future Work
- Try different camera type to distinguish doctor/nurse/etc.
- Use information to detect anomalies

References

Acknowledgement
Thank you to Stanford Health Center and Stanford Artificial Intelligence Vision Lab for the dataset.