Depth-Based Activity Recognition with Convolutional Neural Networks

Alisha Rege, Sanyam Mehra, Alyssa Vann
{ amr6114, sanyam, avann } @ stanford.edu

SVL HEALTH

STANFORD

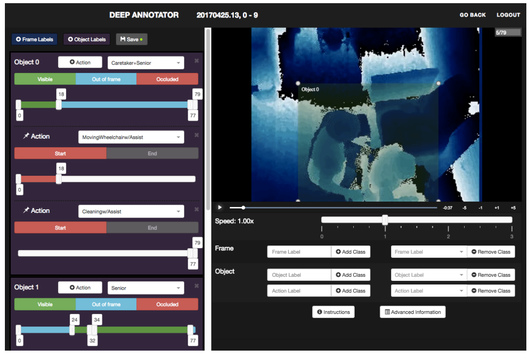
Purpose

- Use computer vision techniques to build an integrated solution for the remote monitoring, assessment, and support of seniors living independently at home.
- Labelled data from depth sensors in the On Lok Senior Home in San Francisco
- Build two state-of-the-art convolutional neural networks
- Created a data labelling tool

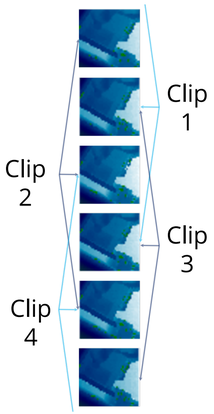
Dataset

- 9 depth sensors/ 4 rooms
- 12K video clips collected
- 28 Activities
- Per frame labels for person, bounding box, activity
- Video clips 7 x 224 x 224 x 1 (frame x width x height x channel)
- Batch size of 4 video clips

Sample depth images



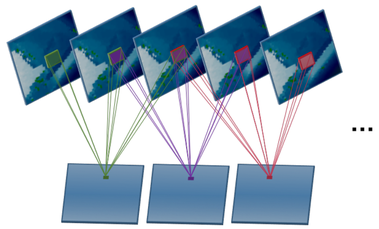
DeepAnnotator Website



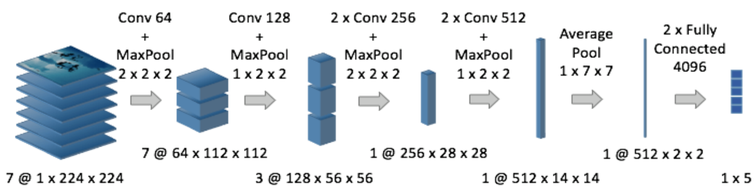
Clip Formulation

3D Convolutional Neural Network

- 3D Convolutional Layers
- Batch Normalization
- ReLU activations
- 3D Max Pooling
- 3D Average Pooling
- Fully Connected Layers



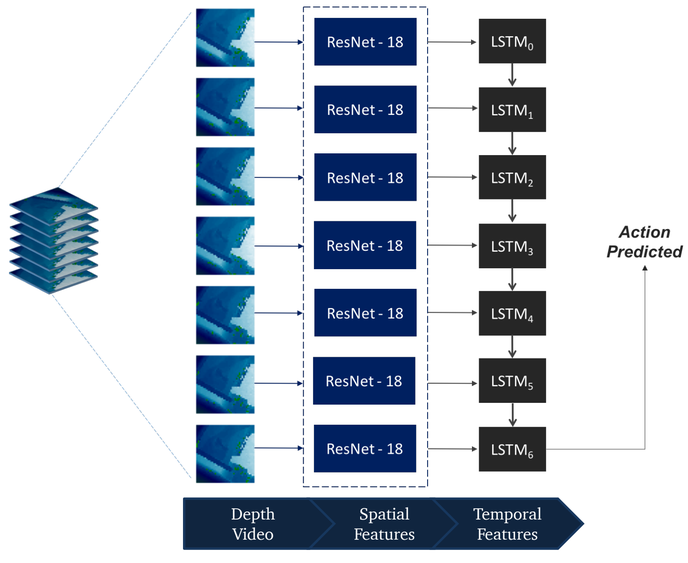
3D convolution



3D Convolutional Neural Network.

Recurrent Convolutional Neural Network

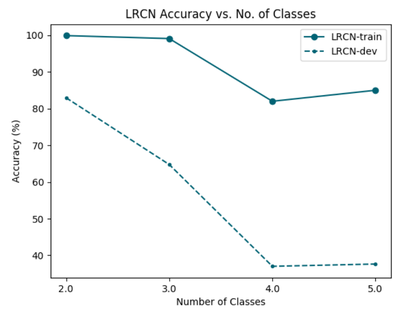
- Pretrained Resnet for spatial feature extraction
- LSTM recurrent network for temporal structure

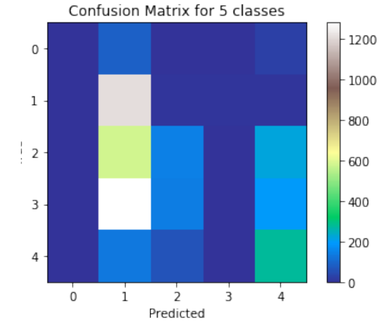


Long-term Recurrent Convolutional Network

Results

Model	Pretrained	LR	Hidden size	Num Layers	Number Of Classes	Best Train Accuracy	Best Dev Accuracy
LRCN	Y	1.00E-04	2000	2	2	66	76
LRCN	Y	5.00E-04	2000	3	3	99.09	64.76
LRCN	Y	1.00E-04	2000	3	4	82.41	37.09
LRCN	N	1.00E-04	1000	2	5	79.8	29
LRCN	Y	1.00E-04	2000	2	5	68.06	28.92
LRCN	Y	1.00E-04	2000	3	5	72.81	32.34
LRCN	Y	1.00E-04	2000	4	5	58.57	21.26
3D Cov	N	1.00 E-03	n/a	n/a	5	99.32	38.23





Future Work

- Record and label more data
- Extend to person localization – from annotations created in dataset
- Activity recognition from point cloud representation
- Feature extraction using optical flow, depth-based descriptors
- Extend to other modalities; Thermal sensor data

References

[1] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4489-4497).