Action Recognition with Depth and Thermal data using Densely Connected Convolutional Network

Vivian Yang
viviany@stanford.edu
Department of Electrical Engineering

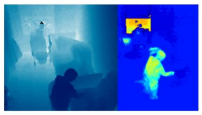
Introduction

Nowadays, the increase of senior population is a worldwide problem that is waiting to be solved and is actually closely related to our life. To improve the living quality of seniors living alone at home or in senior care facilities, we aim to create a vision-based system to monitor seniors' daily behavior. In our paper, we gather our own dataset by installing depth and thermal sensors at senior homes, annotate the collected video clips for specific actions and apply convolutional networks for action recognition. Our classification model can achieve mAP 0.9329 for depth, 0.9485 for thermal, and 0.9544 for the combination of two modalities.

Data

We annotated 7 days of data, which includes a total of 1239 clips with both depth and thermal modality. We focus on 4 fundamental activities: sleeping, sitting, standing, and walking. All other actions are categorized into the background class.

Action	Clips		Frames		Frames per clip	
Sitting	280	107	18287	7314	65	68
Sleeping	33	13	7730	1147	234	88
Standing	181	90	10822	5516	60	61
Walking	129	72	3519	2010	27	28
Background	250	84	36663	13654	147	163
Total	873	366	77021	29641	88	81



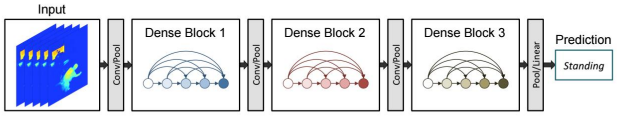
← The statistics of dataset. Left: training data, containing 5 days of videos. Right: test data, containing 2 days of videos.

Conclusion & Future Work

From the result we can see that the performances of depth and thermal are almost the same, both of them achieve about 90% overall accuracy and over 0.94 mAP. Moreover, the result shows that the combination of two modalities has the best results: about 93% overall accuracy and over 0.95 mAP, which proves that the performance improves when modalities are combined together.

In the future, we might extend this task into a temporal action detection task by using our current model with smoothing window or a network that is create for video analysis, such as Convolution3D.

Approach



DenseNet: Densenet is the network that contains shorter connections between layers close to the input and also layers close to the output. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers.

Input: Concatenate L frames of each clip into a L × C channel image, so the size of the input image will be H × W × (L C), where C = 1 in both depth and thermal modalities.

1. Train: Randomly sample clips of length L frames from the training video.
2. Test: Uniformly sample N clips of length L frames from the video and average the results.

Multi-modal Training: Train models for depth and thermal separately, and then during the test time, average the results of the last layer

Experimental Results

	Model	Overall Accuracy	Mean AP
Depth	Dense 121	0.896175	0.948458
	Dense 169	0.855191	0.900562
	Dense 201	0.896175	0.913582
Thermal	Dense 121	0.887978	0.945820
	Dense 169	0.896175	0.932860
	Dense 201	0.904372	0.944655
Combine	Dense 121	**0.926230**	**0.954350**

← With Learning rate = 5e-3, Batch Size = 8, Epoch = 40

↓ Confusion matrices. Left: depth; middle: thermal; right: combined. Index 1 to 5 indicate the categories *sitting*, *sleeping*, *standing*, *walking*, and *background*.