CS230

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

Vivian Yang
Department of Electrical Engineering
Stanford University
viviany@stanford.edu

Abstract

Aging is a big problem in many countries around the world. Right now, 1.3 million people in the US live in senior care facilities. However, the services and manpower at senior homes are usually not enough to support the seniors' daily lives. In this paper, we first collected a real-world action recognition dataset consisting of privacy-preserving depth and thermal signals. Then, we introduce a vision-based monitoring system that classifies daily activities of seniors, which helps facilities to monitor the seniors' daily lives and improve the well-being of seniors.

1 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 work can break into several parts. First, install sensors at senior homes to collect then annotate data. Next, train DenseNet [1] models to perform action recognition on our dataset. Finally, evaluate the performance of each modality and the combination of two modalities, then analyze the advantages and disadvantages of each modality and see if they provide complementary information.

2 Related work

Action classification for RGB images has been studied by a lot of researchers for many years. Tran *et al.* [3] proposed a 3-dimensional convolutional network (C3D) that add on an extra time dimension to preserve the temporal information of the input signals. Donahue *et al.* [4] trained an LSTM layer on top of CNN, where the RNN structure is able to capture the temporal information. The success in RGB video understanding also leads to a series of studies on action recognition with different modalities [5, 6, 7].

3 Dataset and Features

Depth modality. Pixel values in a depth image represent the calibrated distance between the corresponding object and the depth sensor. Comparing with RGB images, depth images have several

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

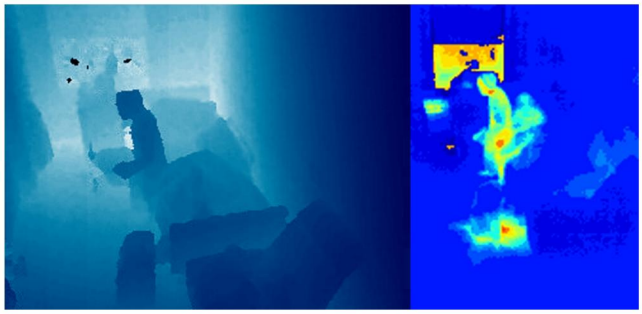

Figure 1: Left: depth image. Right: thermal image.

advantages: work in any lighting conditions, color and texture invariant, can capture the 3D structures of the scene. We use ASUS Xtion PRO as our depth sensors. It records depth images at 240x320 resolution and 30 fps.

Thermal modality. Pixel values in a thermal image represent the temperature of the corresponding object. Similar to depth images, thermal images do not suffer from low light conditions and are also color invariant. We use FLIR Lepton 3 as our thermal sensors. It records thermal images at 160x120 resolution and 8.8 fps.

Camera alignment. The depth and thermal cameras have different frame rates, so the first step is to align them temporally. All the frames are recorded with timestamps, and thus we alignment them by using nearest neighbor matching. For spatial alignment, since the depth and thermal sensors are installed close to each other, they have roughly the same view. Also, our algorithm does not require two camera to be spatially aligned, so it is unnecessary for us to calibrate the images.

Annotation. We annotated 7 days of data, which includes a total of 1239 clips and 106662 frames with both depth and thermal modality. We focus on 4 fundamental activities [2]: *sleeping*, *sitting*, *standing*, and *walking*. All other actions are categorized into the background class. The background class contains a wide variety of videos, including the senior using toilet or changing clothes, anyone other than the senior (e.g. a housekeeper), multiple people in the room (e.g. a caregiver assisting the senior) and empty room.

Train and test split. We split the dataset by date: the first 4 days and the 6th day as training set and the 5th and 7th days as test set. The statistics of the datasets are summarized in Table 1 and 2.

Action	Clips	Frames	Frames per clip
Sitting	280	18287	65
Sleeping	33	7730	234
Standing	181	10822	60
Walking	129	3519	27
Background	250	36663	147
total	873	77021	88

Table 1: Statistics of the training data, containing 5 days of videos.

Action	Clips	Frames	Frames per clip
Sitting	107	7314	68
Sleeping	13	1147	88
Standing	90	5516	61
Walking	72	2010	28
Background	84	13654	163
total	366	29641	81

Table 2: Statistics of the test data, containing 2 days of videos.

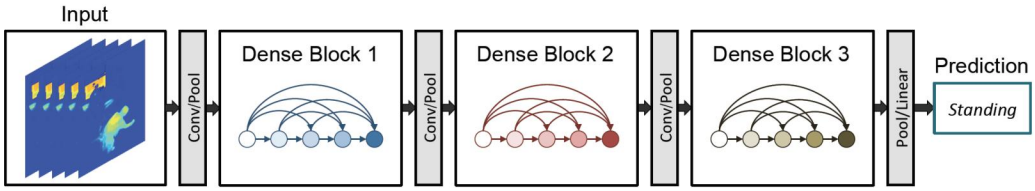

Figure 2: DenseNet architecture.

4 Methods

4.1 Activity Classification

Input. Let a video be $x \in \mathbb{R}^{T \times H \times W \times C}$, where T is the number of frames in the video. Our model $f(x)$ outputs the probabilities of each class, and our goal is to find a f that minimizes the loss of per-clip classification. During training, we randomly sample clips of length L frames from the training video. At test time, we uniformly sample N clips from the video and average the results. In DenseNet models, we concatenate L frames of each clip into a $L \times C$ channel image, so the size of the input image will be $H \times W \times (LC)$, where $C = 1$ in both depth and thermal modalities.

Loss function. The loss function we use is the cross-entropy loss function in Pytorch, which is useful for classification task with C classes. It can be described as:

$$loss(f(x_i), y_i) = -\frac{\log(\exp(f_{y_i}(x_i))}{\sum_j \exp(f_j(x_i))},$$

where $f_j(x_i)$ is the probability score of class j.

4.2 Model

The model architecture we use in this project is Densely Connected Convolutional Network [1]. Densenet is the network that contains shorter connections between layers close to the input and also layers close to the output. As shown in Figure 2, for each dense block, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers.

However, in the original DenseNet, it only takes single image $x \in \mathbb{R}^{H \times W \times C}$, where $C = 3$ as input, yet our input is a video clip $x \in \mathbb{R}^{L \times H \times W \times C}$ that is resize to $x^* \in \mathbb{R}^{H \times W \times (LC)}$. Thus, we need to modify the "conv0" layer to have the correct input channel. Moreover, since we are taking the pretrained parameters to fasten the training process, we will need to transform the original 3 channel weight in "conv0" layer to LC channel, and also remove the weights for the last fully-connected layer.

4.3 Multi-modal Recognition

Since we have two modalities, it is reasonable to combine them together to gain a better performance. One of the methods to fulfill this goal is to since we have two modalities, we can combine them together to gain a better performance. One of the methods to fulfill this goal is to train two separate models for depth and thermal, f_T and f_D, and then during the test time, average the results of the last layer $f(x) = (f_T(x) + f_D(x))/2$.

5 Experiments

Evaluation Function. We calculate the overall accuracy, mean average precision, and draw out the confusion matrix. Overall accuracy is a standard evaluate metric for classification, but mean AP can better deal with imbalanced dataset. To show out the strength and weakness of the model, we include the confusion matrix as well.

Hyperparameter Tuning. From Table 3 and 4 we can see that DenseNet 121 with learning rate = 5e-3 performances best for both depth and thermal modalities. Therefore, we pick out these two models to look at their output details and do the multi-modal recognition by using their outputs of the last layer.

Model	Learning rate	Accuracy	Mean AP
DenseNet 121	1e-3	0.8852	0.8909
	5e-3	**0.8962**	**0.9485**
	1e-2	0.8634	0.8981
DenseNet 169	1e-3	0.8798	0.9158
	5e-3	0.8552	0.9006
	1e-2	0.8306	0.8777
DenseNet 201	1e-3	0.8852	0.9156
	5e-3	0.8962	0.9136
	1e-2	0.8443	0.8612

Table 3: Hyperparameter tuning of depth modality.

Model	Learning rate	Accuracy	Mean AP
DenseNet 121	1e-3	0.7978	0.8774
	5e-3	**0.8880**	**0.9458**
	1e-2	0.8169	0.9049
DenseNet 169	1e-3	0.8361	0.9044
	5e-3	0.8962	0.9329
	1e-2	0.8716	0.9105
DenseNet 201	1e-3	0.8197	0.9063
	5e-3	0.9044	0.9447
	1e-2	0.8060	0.9167

Table 4: Hyperparameter tuning of depth modality.

Results. From Table 5 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. Also, from the confusion matrices in Figure 4 we can see that compare to depth model, thermal model performances equally well or better in most of the classes, yet very bad in the walking class. The thermal model always misclassified walking as standing or background. However, the combination of two modalities can preserve or even outperform the best accuracy of each class in depth and thermal models.

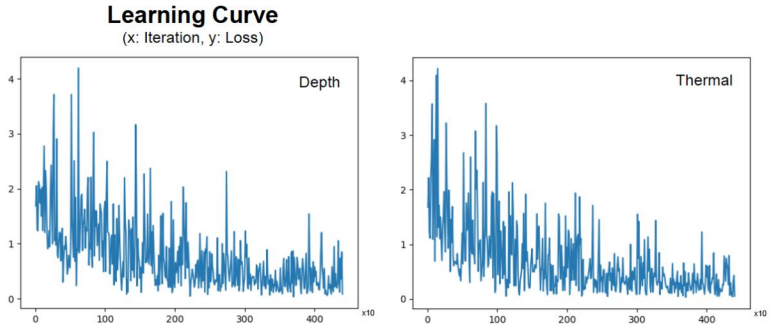

Figure 3: Learning curve of DenseNet 121 with learning rate = 5e-3, batch size = 8, epoch = 40.

Modality	Accuracy	Mean AP
Depth	0.8962	0.9485
Thermal	0.8880	0.9458
Depth + Thermal	**0.9262**	**0.9544**

Table 5: Results of action classification on depth, thermal, and the two combined with DenseNet 121 and learning rate = 5e-3.

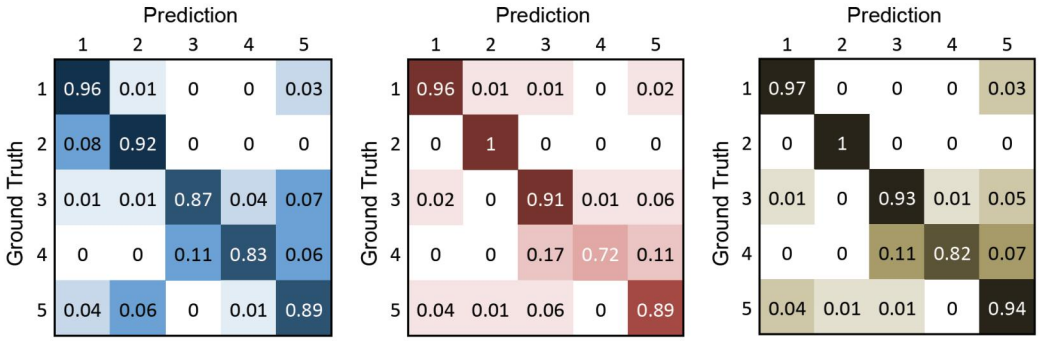

Figure 4: Confusion matrices. Left: depth; middle: thermal; right: combined. Index 1 to 5 indicate the categories sitting, sleeping, standing, walking, and background.

6 Conclusion

This paper introduces a vision monitoring system based on privacy-preserving signals that accurately monitors the fundamental daily activities of seniors. We also introduced a newly collected real-world action classification dataset with both depth and thermal modalities, and the performance of our action classification model is great.

In the future, we will try to 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 C3D [3] mentioned in related work.

References

[1] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] E. A. Phelan, B. Williams, B. W. Penninx, J. P. LoGerfo, and S. G. Leveille. Activities of daily living function and disability in older adults in a randomized trial of the health enhancement program. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(8):M838– M843, 2004.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. *International Conference on Computer Vision (ICCV)*, 2015.

[4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[6] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. *AAAI workshop on Pattern, Activity and Intent Recognition*, 2011.

[7] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.