Multiple Sensor Fusion for Human Activity Recognition

Ronaldson Bellande
University of Massachusetts Lowell
Dept of Computer Science
Ronaldson_Bellande@student.uml.edu

Ashwin Jagadeesha
University of Massachusetts Lowell
Dept of Computer Science
ashwin_jagadeesha@student.uml.edu

Jay Pandya
University of Massachusetts Lowell
Dept of Mechanical Engineering
JayAtulbhai_Pandya@student.uml.edu

***Abstract*—Deriving activities of humans from a video feed is a laborious activity and directly falls into the domain of computer vision but is susceptible to noise from various sources. Although the methods are accurate with the surge in current computer vision and computing capabilities there is a scope for improvement in the activity recognition. While, tagging the actions being performed by a human from a video feed with high accuracy is a quality of an intelligent AI based activity recognition system. However, in safety critical applications it is imperative to utilize multiple sensor modalities for robust operation. To exploit the benefits of state-of-the-art machine learning techniques for Activity Recognition (AR), it is valuable to have multi-modal data-sets.**

Using camera to determine what Interaction needs to be recognize the environment and the human activity. In general, activity recognition is to identify the set of actions and objectives of one or more objects from a series of examination on the action of object and their environmental condition. The major applications of Human Activity Recognition vary from Content-based Video Analytics, Robotics, Human-Computer Interaction, Human fall detection, Ambient Intelligence, Visual Surveillance, Video Indexing and so on.

In this project we also aim to examine the challenges, and issues of Human Activity Recognition systems in a sensor fusion based model. There are a lot of variants in Human Activity Recognition systems such as Human Object Interactions and Human-Human Interactions etc. We intend to create a data-set which collects different activities from this sensor fusion and study their properties whose findings will be used to create an accurate mapping from data using Machine Learning and Deep Learning to determine what Human activity is being performed.

I. Introduction

Human Activity Recognition gains its popularity due to its versatility. The goal of human activity recognition is to identify a set of actions performed by humans in an unobtrusive manner from a series of examinations to draw their actions into an activity. As a single activity corresponds to many elementary actions, applications of this system are not limited to surveillance security, sports, and context-based inferences for each activity.

A very important reason for utilizing the outcome of an activity recognition model is to use them in human healthcare applications and human computer interaction based applications in a very effective manner.

Fig. 1. Acoustic environment activity example

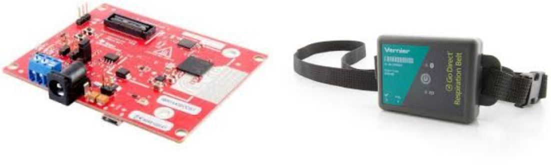

Fig. 2. IWR1443 radar and Vernier GoDirect Respiration belt

Traditional activity recognition models have performed really well with highly accurate classification but due to their limitations such as availability of sensors and Fig. **??** such as hospitals or other open acoustic environments there is a demand for innovative sensor fusion based models which could take advantage of such environments Fig. 3.

So the next step was to select sensor which specifically targets Acoustic environments such as in Fig. 3 and create a fusion of sensors for activity recognition which motivates us to use 2 sensors one being IWR1443 ER from Texas instruments Fig. 2 and the other one being Vernier Go Direct respiration belt Fig. 2

The next step is data acquisition from these 2 sensors to create a multi-modal sensor fusion to perform activity classification Fig. 5.

A total of 9000 samples of human activity data for walking, standing and sitting were collected using Respiration belt and a total of 6800 samples which were further up-sampled to 9000 samples. The up sampling technique involved random sampling from the same data-set. A total of 9000 samples consisting of 1D force data over time were collected from Respiration belt. Initial data prepossessing was done to identify the normal distributions of activity data. Results in Fig. **??** below shows clear indication of activities showing variance. Increase in the force value from static activity of sitting and standing

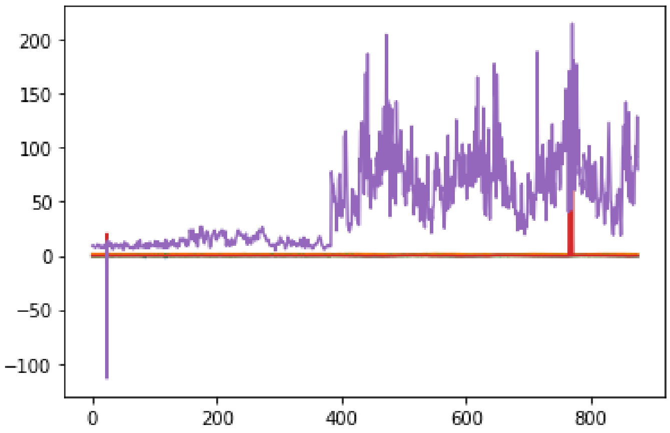

Fig. 3. Heart rate distribution over 9 minutes

to dynamic activity of walking improves confidence on sensor reading for building models for activity classification.

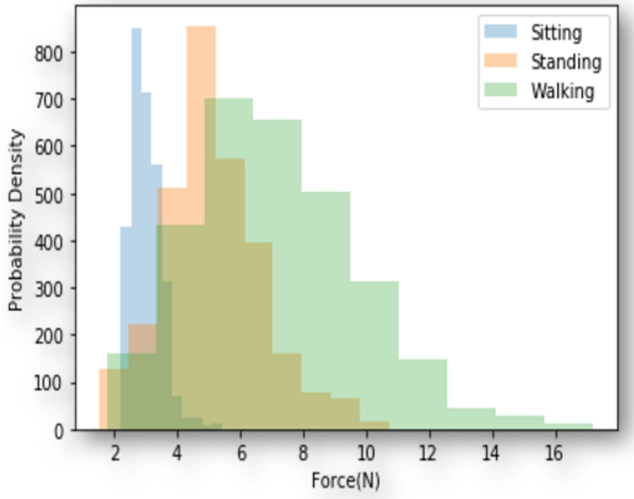

Fig. 4. Histogram of Activities

Goal was to detect activities such as sitting, standing and walking by building a deep neural network that learns from features from millimeter wave and vernier respiration belt sensor Fig. 6. And trains itself to predict activities accurately. And this forms the base for a combined model which is discussed in the greater detail in section 2.

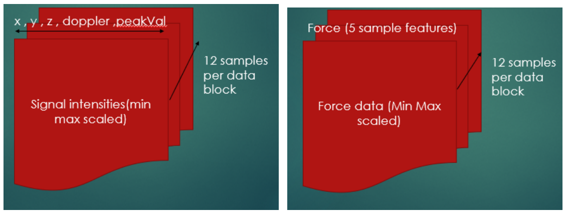
x , y , z , doppler ,peakVal
12 samples per data block
Signal intensities(min max scaled)
Force (5 sample features)
12 samples per data block
Force data (Min Max scaled)

Fig. 5. Data structures of mmWave and Respiration belt

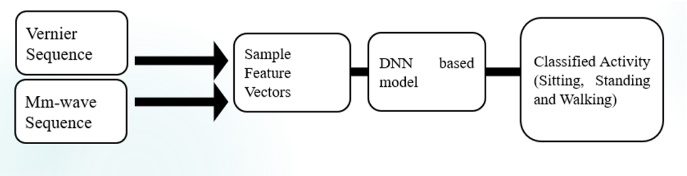

Fig. 6. Workflow for classification

II. Methods

The approach we took to build a fusion model was to black box the deep-learning approach initially to try different DNN architectures and then start bench marking each of these approaches based on performance, once we were able to zero in on the right DNN architecture we worked on fine tuning the architecture to yield better results for classification. Then, we observed that the models were stable performance wise so we then moved towards a fused DNN based classification models into one single DNN architecture for activity classification. So we started our experiment by selecting LSTM first and and then moved onto 1D CNN based models to see if the performance improved as both these signals had typographical connections in their signals hence LSTM and CNN were selected which works well on such data.

A. Long Short-Term Memory

The first approach we took was to build 2 individual LSTM models one for classification of activities from millimeter wave data and another LSTM model for Respiration belt data to draw some performance metrics for these models.

The reason for using an LSTM based network is because LSTM is an is an artifitial recurrent neural network(RNN) that is in the field of deep learning that has feedback connections. Sequence classification is a predictive modeling technique where the model is given a sequence of inputs over space or time and the task is to predict a category for the sequence. For our model force is the input sampled at one tenth of a second and mmWave data is sampled at 1 eighth of a second. Hence, we trained these different sensor data individually on 2 different LSTM based networks to draw accuracies.

An LSTM network computes a mapping from an input sequence x =$(x_1, ..., x_T)$ to an output sequence y =$(y_1, ..., y_T)$ by calculating the network unit activation's using the following figure 7 showing equations iteratively from t = 1 to T:

where the W terms denote weight matrices (e.g. W_{ix} is the matrix of weights from the input gate to the input), W_{ic}, W_{fc}, W_{oc} are diagonal weight matrices for peephole connections, the b terms denote bias vectors (b_i is the input gate bias vector), is the logistic sigmoid function, and i, f, o and c are respectively the input gate, forget gate, output gate and cell activation vectors, all of which are the same size as the cell output activation vector m, . is the element-wise product of the vectors, g and h are the cell input and cell output activation functions,[9] generally and in this paper tanh, and ϕ is the network output activation function, softmax in this paper.

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f) \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o) \tag{4}$$

$$m_t = o_t \odot h(c_t) \tag{5}$$

$$y_t = \phi(W_{ym}m_t + b_y) \tag{6}$$

Fig. 7. Working of LSTM model

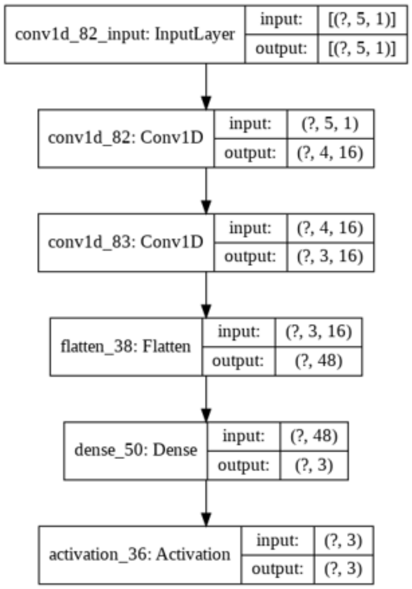

Fig. 8. 1D CNN architecture

B. Convolutional Neural Network

Convolutional neural network was also used to train the data and check the performance. [10] The convolution for one pixel in the next layer is calculated according to the formula given below.

net (i, j) = (x * w) [i, j] = $\Sigma_m \Sigma_n x[m, n]w[i - m, j - n]$

Where net (i, j) is the output in the next layer, x is the input image and w is the kernel or filter matrix and * is the convolution operation. As can be seen, the element-by-element product of the input and kernel is aggregated, and then represents the corresponding point in the next layer.

In this part we modeled the data into a 1D sample of 1800 data-points for both millimeter wave and Respiration belt to have the same dimensional and sampling method. And then we trained each of these data-sets on 2 different 1 dimensional CNN based networks. The network architecture used for this network is Fig. 8.

C. Multi input single output 1D CNN

In this part we modeled the data into a 2 input network with each taking in a 1 dimensional feature vector with a total sample size being 1800 data-points split into training and validation with the split being 8:2 for both millimeter wave and Respiration belt to have the same dimensions and class label columns. The loss function that is employed is categorical cross entropy based loss. And then we trained each of these data-sets on 2 different 1 dimensional CNN based networks. The network architecture used for this network is Fig. 9. The results for this network are discussed in the Results section of this paper.

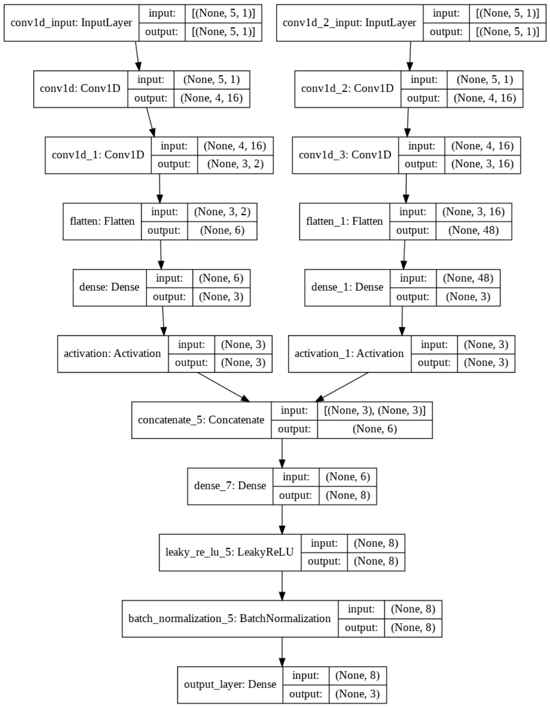

Fig. 9. Multi-input 1D CNN architecture

III. Dataset

There were three data-sets two of them were collected from 2 different devices for 3 activities (such as walking, sitting and standing) and the other data-set was required that has 5 activities and 3 of them were identical activities to the collected data-sets. The data collected by the mm-Wave device and vernier device were collected at the same time for specific activities but due to the delay the time frequencies are not in sync.

A. Activity Data set collected

Collecting data from mmwave device and vernier device for activities such as standing, sitting and walking. The collected data for the mmwave for activities had little errors specially when it came to an activity that would relatively be the same but in a much slower passed like, specially for like standing and sitting. The walking error prediction for the training data-set was 37% and for the testing data-set it was 42%. The standing error prediction for the training data-set was 24% and for the testing data set it was 19%. The sitting error prediction for the training data-set was 24% and for the testing data-set it was 23%. The walking prediction error for walking was was

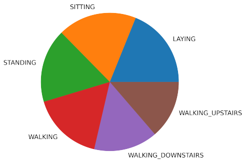

Fig. 10. Five activities chart diversity

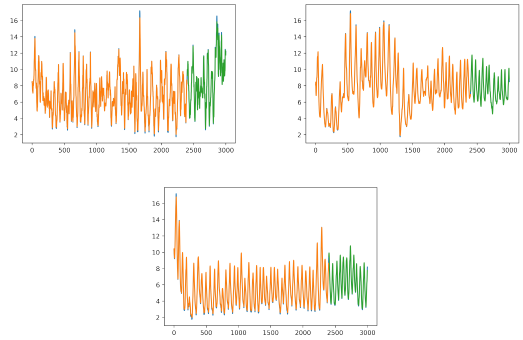

Fig. 11. Activities error for walking, standing and sitting

high for the training data-set and testing data-set than any other activity since it had a wide array of range of that the walking can be since sitting and standing was activities that you need to not be moving and for walking you can be walking fast or walking slow.

B. Activity Data set by Related Works

Data-set was collected by related works for activity recognition, the activity that the data set has are walking, standing, sitting, laying, walking upstairs and walking downstairs. The data-set has an equal amount of data for each activities so there is no activity that has more importance while training.

Activity of collected data error prediction for training
Walking 37%
Standing 24%
Sitting 24%

TABLE I
ERROR PREDICTION FOR THE TRAINING DATA SET

IV. Results and Discussion

Our 1D CNN based model trained on millimeter wave achieved around 93.44% accuracy on validation and another 1D CNN model trained on Respiration belt achieved around 82.13% accuracy in predictions on validation data. The combined 1D CNN model trained on a 2 input 1D CNN based model achieved a validation accuracy of around 97.3%.

V. Conclusions

The combined activity recognition model works well with the right procedures for preparing data and the use of either deep learning or any other ML algorithm. This also provides a lot of possibilities of being able to predict classification outcomes in the absence of sensor modalities in a multi-sensory network if there is a failure or simply absent. Mm-Wave waves versatility allows to model more robust approaches to sensing and classification activities in multiple domains which involves movement.

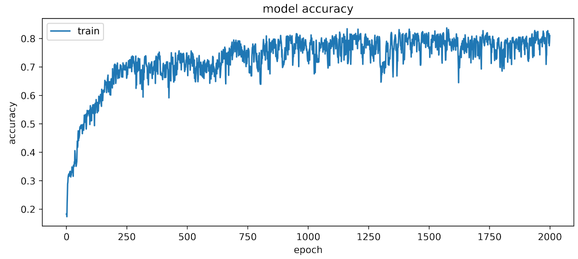

Fig. 12. Model accuracy for the 5 activities converging

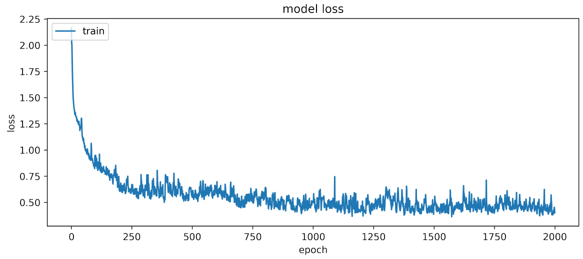

Fig. 13. Model loss for the 5 activities converging

VI. Future Work

The next step would be to build a Multitask learning based deep neural network that predicts breathing rate of a human subject in the absence of data from Respiration belt subject to activity classification which is only based on millimeter Wave

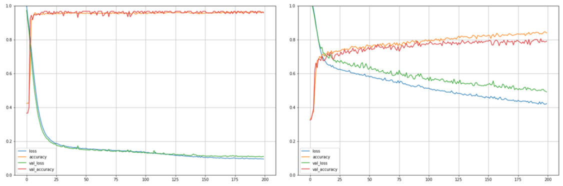

Fig. 14. Valdiation loss and accuracy from individualt 1D CNN models for mmwave and Respiration belt data

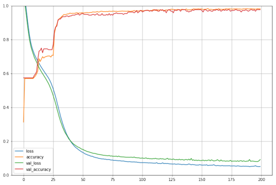

Fig. 15. Valdiation loss and accuracy from multi-input 1D CNN model for mmwave and Respiration belt data

Model	Training loss	Validation loss	Training Accuracy	Validation Accuracy
1D CNN mmwave	0.177	0.142	94.33%	93.12%
1D CNN vernier	0.162	0.172	87.69%	84.33%

TABLE II

METRICS FOR 1D CNN BASED TRAINING OF MODELS

Model	Training loss	Validation loss	Training Accuracy	Validation Accuracy
1CNN Combined	0.018	0.022	96.33%	97.12%

TABLE III

METRICS FOR 1D CNN BASED TRAINING OF MODELS

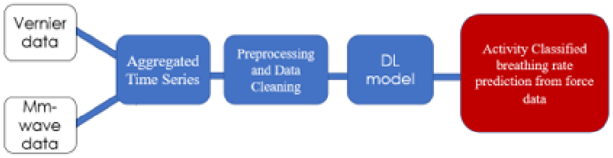
Vernier data
Mm-wave data
Aggregated Time Series
Preprocessing and Data Cleaning
DL model
Activity Classified breathing rate prediction from force data

Fig. 16. Future Process Layout

data. With the help transfer learning to help merge models to help predict human activities from a noninvasive fashion which also predicts the breathing rate from a well trained breathing rate model (without using vernier force sensor) as seen in the Fig. 16.

REFERENCES

[1] Mirco Möncks, Varuna De Silva, Jamie Roche, and Ahmet Kondoz, “Adaptive Feature Processing for Robust Human Activity Recognition on Novel Multi-Modal Dataset.

[2] Csaba Benedek, Bence G’alai, Bal´azs Nagy and Zsolt Jank´o, “Lidar-based Gait Analysis and Activity Recognition in a 4D Surveillance System.

[3J K. Aggarwal, L. Xia, O. C. Ann, and L. B. Theng, “Human activity recognition: A review,” Pattern Recognit. Lett., no. November, pp. 28–30, 2014

[4] L. Pishchulin et al., “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation,” in 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] P. Kumari, L. Mathew, and P. Syal, “Increasing trend of wearables and multimodal interface for human activity monitoring: A review,” Biosensors and Bioelectronics, vol. 90. pp. 298–307, 2017.

[6] A. Jalal, Y. H. Kim, Y. J. Kim, S. Kamal, and D. Kim, “Robust human activity recognition from depth video using spatiotemporal multi-fused features,” Pattern Recognit., vol. 61, pp. 295–308, 2017.

[7] J. Zhu, R. San-Segundo, and J. M. Pardo, “Feature extraction for robust physical activity recognition,” Human-centric Comput. Inf. Sci., vol. 7, pp. 1–16, 2017.

[8] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, “A Review on Human Activity Recognition Using Vision-Based Method,” Journal of Healthcare Engineering, vol. 2017. 2017.

[9] Has¸im Sak, Andrew Senior, Franc¸oise Beaufays, ”Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling”

[10] Saad Albawi , Tareq Abed Mohammed ,”Understanding of a Convolutional Neural Network”