Human Activity in Autonomous Vehicles

Mohammad Ershad Shaik
UT Austin, USA
mohammad.ershad@utexas.edu

Chongyan Chen
UT Austin, USA
chongyanchen_hci@utexas.edu

Tim Yao
UT Austin, USA
tyao@utexas.edu

ABSTRACT
This paper presents a novel method of driver activity recognition for autonomous vehicles. We identify the key activities that a driver will perform in an autonomous vehicle and collect recording of these activities. We then build a predictive model to recognize those activities with training data from 8 subjects. Our generalized model have achieved up to 84% accuracy using LSTM.

Author Keywords
LSTM, Pose Estimator, Autonomous Vehicle, Driver Activity Recognition

INTRODUCTION
Recognizing human activities is crucial and widely needed in smart homes, health care, and so on. Our work aims to recognize human activities in self-driving cars to address the challenge with take-over requests and allow manufacturers to build a more human centric vehicle. As conditional autonomous driving gains more traction in the near future, especially in the field of self driving trucks, the vehicle must understand what the driver is doing to safely and timely request a take-over [17,10]. To perform this activity recognition, we propose a two stage ensemble that uses a pose estimator to extract the driver's skeleton model and a random forest or LSTM to perform classification. The feature extraction method using the pose estimator reduces the amount of data required, is robust to all drivers and vehicles, and protects the privacy of the driver (as no images need to be stored). For classification, general model using a LSTM and a random forest classifier.

PRIOR WORK
Previous studies [1,2] have been conducted to understand driver behaviours in autonomous vehicles. Of these, researchers have shown a dependence of take-over time and quality on driver's activity and distraction level [11,12] Other studies have explored traveller preferences in AVs [3] compared to existing mode of travel. Most of these studies are based on surveys conducted after several hours of driving or through a virtual simulation environment. Additional studies [4,5] also show that drivers are more likely to perform "household activities" in AVs. While these prior research place a heavy emphasis on the investigating the activities performed in AVs and their potential impact on the vehicle decisions, none of these studies use prediction models to predict driver activity using visual data processing.

Recently however, researchers have started to perform driver activity recognition for autonomous vehicles. In 2015, researchers used eye and hand tracking to predict whether the driver is performing one of the four activities [15]. This achieved an acceptable overall accuracy of 77% for 4 activities (idle, video, read, and email) . Very recently, an ongoing project from MIT [13] is attempting to build a human centric autonomous vehicle using a deep learning approach [14]. Part of this project is detecting the driver's activity. While no accuracy results have been published yet, their method involves using gaze tracking and body movement tracking through optical flow.

In the area of general activity recognition through pose estimation, previous research has also shown high promise. The models used in these research are similar to our two stage ensemble model [16]. A pose estimator performs initial feature extraction and then a second stage performs the classification. For the second stage, various models have been used including an ensemble of SVMs (90% accuracy) and a RNN (94.5% accuracy) [17].

ACTIVITY IDENTIFICATION
We narrowed down the activities to the domain that is "commuting from home to work" of a working professional. Literature [1-7] suggests that autonomous vehicles (AV) will enable drivers to catch up on the activities like media watching, work, reading, and sleep they are sacrificing for long travel times. To investigate this, we collected one subject data commuting from home to work on "non self driving car" and observed that subject performed different activities while stopped at traffic signals such as eating, using cell for navigation, talking, messaging on cell, listening to media, and stretching. We think that drivers will perform these "stopped" activities more frequently when AVs become widely available. Based on this we have decided to focus on recognizing and classifying 5 major activities of: eating, talking on cell, messaging on cell, reading a book, and a baseline of hands on steering wheel in a stationary car.

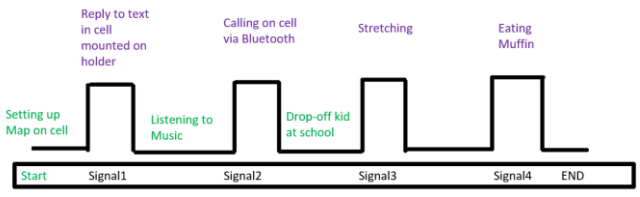

Figure 1. Subject activity in a moving car during stop at signals while commuting from home to work on a routine day.

DATA COLLECTION AND ANNOTATION

We collected data on a total 10 subjects. Only 8 subjects were used as the other two subjects had poor video quality that will negatively impact the accuracy. Subjects were asked to sit in the driver's seat of a stationary car. An iPhone 6s with a wide angle lens is mounted on top of the front drivers glass with good visibility to steering wheel and subject for capturing videos as shown in Figure 2.

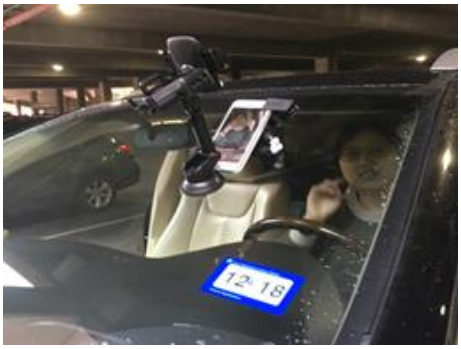

Figure 2. The camera mounting and positioning for recording "drivers" performing activities.

The subject is then asked to "text message" on cell for 1 min and then with a voice command, the subject is asked to take over control with "hands on steering wheel position (baseline)". After 30 sec subject is asked to "make a phone call" for 1 min with switching of the cell phone from left to right hand. The subject is then asked to place "hands on wheel" for another 30 sec. Next, subject is asked to "read a book" for 1 min and then switch back to hands on steering wheel for 30 sec. Next, subject is asked to "eat some food" for 1 min and finally coming back to steering wheel for 30 sec. 4 different human activities are performed with a baseline "hands on steering" position interleaved in between each activity. All these videos are captured with 1080P 30fps resolution and are processed in the later steps. Table 1. shows different activities that subjects performed, start time for each activity, end time for each activity with intervals tabulated.

Start (min:sec)	End	Activity	Interval (sec)
0	30	Steering	30
0:30	1:30	Texting	60
1:30	2:00	Steering	30
2:00	2:30	Calling Right hand	30
2:30	3:00	Calling Left Hand	30
3:00	3:30	Seering	30
3:30	4:30	Reading	60
4:30	5:00	Steering	30
5:00	6:00	Eating	60
6:00	6:30	Steering	30

Table1. Data collected from each subject with time interval for each activity.

For data annotation, we made special accommodations for the calling and eating activities. For calling, we specifically labelled whether the calling was done with the left or right hand. For eating, we annotated the motions of bring food to and from the mouth. As a result, we have a total of 7 subclasses: steering, texting, calling left, calling right, reading, eating to mouth, and eating to lap. As will be discussed in the model design section, this provides an accuracy improvement over naive annotation of the target 5 activity classes.

CONTRIBUTIONS AND APPROACH

Unlike previous research in the field of driver activity recognition, we chose a purely visual approach based on driver pose. We believe that such a design can be better generalized to different drivers and large number of activities. Our classifier uses a two stage ensemble network with a human pose estimator as the first stage and a LSTM or random forest classifier as the second stage. This design is largely influenced by our dataset size and model interpretability. From related research, we know that an ensemble network can outperform a pure vision based CNN in activity recognition tasks [16]. However, we do not have enough diverse data to fully train a vision CNN needed for the first stage, therefore, we chose to incorporate a pre trained pose estimation network instead. The second stage classifier, either a LSTM or random forest, then uses these pose information to predict the driver activity.

POSE ESTIMATOR DESIGN

For pose estimation, we tried various models including Posenet and Openpose. We chose a Tensorflow port of the open source model Openpose [6,7,8,9], which is developed from CMU, for its ease of integration and relatively high accuracy. This neural network is pretrained to detect various joints on the human body and output their X and Y locations in the image as shown in Figure 3. For our model, we only use joints located in the upper torso, specifically the wrists, elbows, shoulders, and nose. To reduce inaccuracy noise of the pose estimator, we perform a temporal moving average filter across 5 frames to smooth out any jitters in the pose. Additionally, we use the relative distance of joints to the neck to remove camera and human position biases.

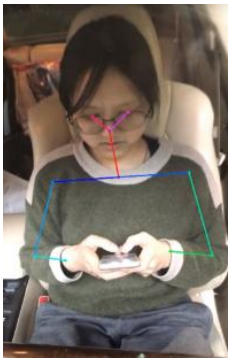

Figure 3: Visualization of the pose estimation

Given that this network is designed to work on all people in all environments, it should be robust for all people (sex, age, race, etc) and all vehicles. A benefit of using pose estimation is that we can preserve the privacy of the driver. Instead of storing videos for training the network, we can only store the poses instead.

SECONDARY FEATURES

Using only the raw pose estimator features gave poor accuracies for texting, reading, and eating, especially with the random forest classifier. In the case of texting and reading, the overall poses are very similar, with the main difference being the distance between left and right hands. Generally, the distance is smaller for phones, but larger for books/magazines. Therefore, we added a secondary feature that computes the distance between left and right wrist joints. Eating detection is particularly challenging as the key distinction between this and other activities is the brief motion of bringing food up to the mouth. This motion is best represented by the right wrist velocity and the distance between right wrist to the nose. Therefore, we created secondary features through these representations. Additionally, as stated in the data collection section, we extracted the motions of bringing food to and from the mouth by explicitly annotating them. Due to the sparsity of these motions, we also perform data augmentation specifically for these eating motions. This is done through generating features based on a Gaussian distribution of the recorded real data. With this, we were able to significantly improve the accuracy of detecting the eating activity. Note that the velocity feature is only used with for the random forest model as the LSTM model already has the ability to understand temporal relationships.

RANDOM FOREST DESIGN

The random forest model uses 100 estimators to predict the 7 subclasses. To capture temporal information, we perform windowing using window sizes of 2 seconds and an overlap of 1 second. These windows are also what allow us to extract the velocities.

LSTM FRAMEWORK DESIGN

The pose estimator extracts the spatial information of the driver activity, so we chose Long Short Term Memory (LSTM) the LSTM network to extract temporal information. By adding a cell state based on RNN, LSTM is good at dealing with time series problem and can address two issues of standard RNN: exploding gradients and vanishing gradients.

The whole LSTM framework has three layers. Two LSTM layers and one Dense layer with “Softmax” activation. The input is a 3D matrix (n samples, 11 time steps, 13 features). Since we have added two distances features that calculated by ourselves, we normalize the features at the very beginning. Also, we have shuffled the data. The train validation split is 0.8/0.2 and the batch size is 1024. The first LSTM layer has 128 outputs while the second LSTM layer has 256 outputs.For the two LSTM layers, we used 0.2 dropout and 0.2 recurrent dropout to prevent over-fitting. For Dense layer, we used softmax activation to give out the possibility of the predicted classes.

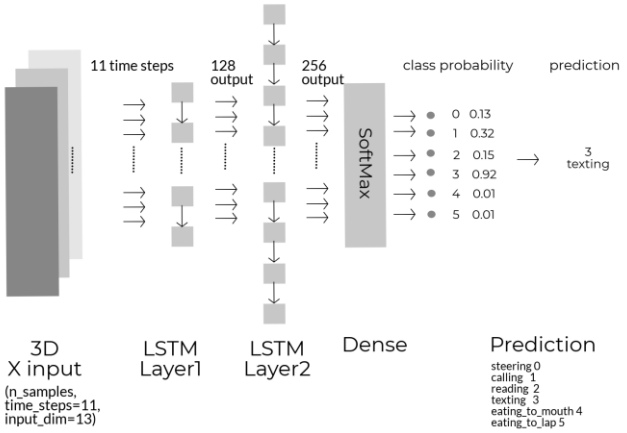

Figure 4. LSTM framework used to predict driver activities

For optimizer, we have compared Stochastic gradient, Adam, and NAdam. We chose Nesterov Adam as our optimizer with a schedule learning rate decay of 0.004. With NAdam, our model can converge quickly with a momentum as shown in Figure 5. The X-axis is epoch while Y-axis is the percentage of loss. The blue line is training loss and the orange line is validation loss. At the beginning, we had a total 1000 epochs, but after nearly 500 epochs, the validation loss stopped decayed, thus we applied early stopping to monitor the validation loss. If the validation loss stops decaying, the training is terminated.

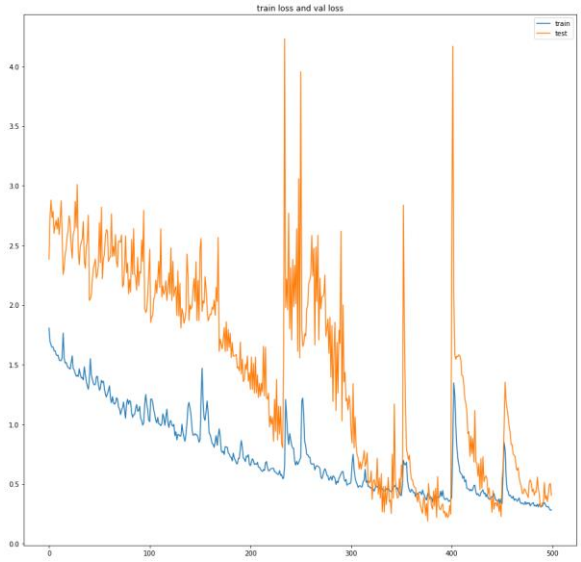

Figure 5. Plot showing entropy loss for training and validation.

RESULTS

We were able to achieve an overall F1 score of 79.1% for the random forest classifier and 84.1% for the LSTM classifier. This is done using a leave one subject out validation strategy. While the overall accuracy is fairly good, we can see that there is a high variance in the accuracy of different classes. Specifically, the reading and eating class accuracies are relatively low as seen in Figure 6 and 7. This is likely due to the inaccuracies of the pose estimator, especially for reading. In many of the reading video data, there are moments when the book/magazine occludes much of the arms and causes the pose estimator to produce incorrect pose estimations. For eating, there is also high variability between subjects in their motions. Some people brought their food all the way down to the lap area while others kept their food close to the mouth.

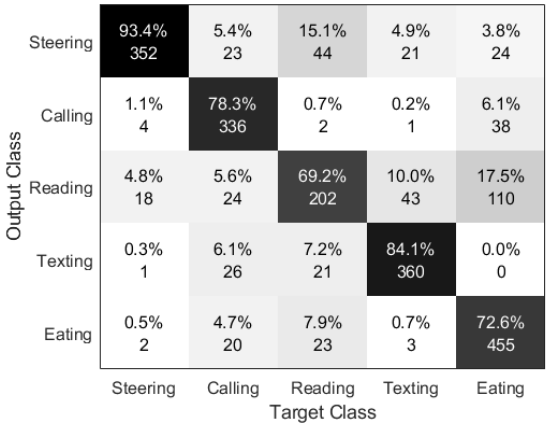

Figure 6. Confusion matrix using random forest classifier

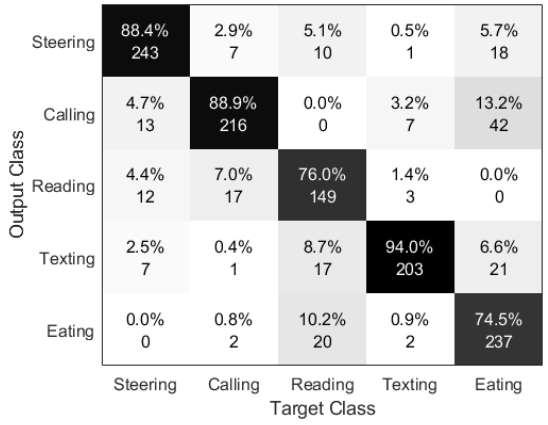

Figure 7. Confusion matrix using LSTM classifier

We also noticed that the LSTM classifier tends to generalize better to all subjects, with the lowest subject accuracy being 57%. The random forest classifier can achieve higher accuracy for some subjects but also much lower accuracy for others. This is likely a sign of overfitting. We believe that for the LSTM, we can achieve even higher accuracy and less inter-subject variability if we had more data. For subject 6, the video quality is relatively poor due to glare on the windshield and the pose estimate was highly inaccurate for the eating and calling activities.

Subject	F1 Score	
	RF	LSTM
1	74%	71%
2	93%	88%
3	52%	72%
4	74%	82%
5	79%	68%
6	44%	57%
7	98%	81%
8	81%	87%

Table 2. F1 scores for each subject.

DISCUSSION

The results show that it is feasible to use a two stage ensemble with a pose estimator and LSTM to perform driver activity recognition. However, the accuracy greatly depends on the pose estimator accuracy. In cases where parts of the body are occluded, the pose estimator will give inaccurate pose estimates. This can corrupt the training of the model and thereby reducing accuracy. One way to alleviate this issue is to use multi-camera pose estimation, but the computational costs will be significantly increased. Additionally, a 3D pose estimator can also increase accuracy while removing camera positioning dependency compared to the 2D estimator we are using currently.

For the LSTM model, we also observed that it requires a large amount of data to properly discover feature relationships and generalize. During testing, a LSTM trained on 3 subjects was completely unable to perform accurate prediction on a 4th subject. Training with 7 subjects however gave us acceptable accuracies. Given that the LSTM achieved higher overall accuracy and lower accuracy variances than the random forest classifier, we can interpret that the LSTM generalizes better.

In hopes to boost accuracy, we also tried to implement an object detection network into the first stage. We used the YOLOv3 network for its high speed to accuracy ratio. While the network can accurately detect the cellphone, it struggles to detect books and food. In many cases, it will misclassify a book as a cellphone. Given this, we chose not to integrate YOLOv3 as it is not accurate enough for our purposes. However, we believe that the accuracy can be improved if we retrained YOLOv3 with only the classes of interest (cellphone, book, various food, etc) instead of the default 80 general object classes.

FUTURE WORK

Camera position and quality have great impact on activity recognition. Currently, we deploy the camera outside of cars' window, which is not practical. For future work, we hope to deploy the camera inside the car using a wider angle lens. In addition, 10 people's data are not enough to fully leverage the LSTM, so more data can potentially give better results as the LSTM learns to generalize better. We also wish to increase the number of activities in the study to include sleeping, putting on face make-up, playing games, watching movie, and conversation with passenger to be collected in an AV simulator. Recently, many unsupervised methods has been proposed that can be used with online learning to predict activities that were not trained on before. Similarly, we can also perform model personalization to increase accuracy for a particular driver. As discussed earlier, switching to a 3D pose estimator and implementing a retrained YOLO can bring higher accuracy and better robustness. We could also improve accuracy by using multi-modal approaches using audio. Lastly, an extension to our model is to also predict take over time from autonomous mode to manual mode.

CONCLUSION

We have identified different driver activities that are relevant for AVs and presented a model that can accurately predict the driver's activity. Using video recordings of various driver activities that we collected from a stationary car, we extracted features using a pose estimator and built prediction models to classify these activities with accuracies of up to 84.1% with a LSTM classifier. We also compared the performance of a RF to LSTM to show the benefits of a LSTM classifier. Given that our model does not require any specialized sensing hardware other than a camera, our design can be economically implemented in all vehicles.

DIVISION OF WORK

Ershad worked on DOE design, data collection, and annotation. Tim worked on Pose estimation, and feature extraction. Chen worked on Random Forest and LSTM model design and Training. Authors have equal contribution to the paper. All of our work can be found here: https://github.com/CCYChongyanChen/Activity-Recognition/

ACKNOWLEDGEMENTS

We thank all the subjects who volunteered for data collection. We also thank Professor Edison Thomaz and Sarnab Bhattacharya for the advice and support throughout the project. .

REFERENCES

1. Saptarshi Das et.al. "Impacts of Autonomous Vehicles on Consumers Time-Use Patterns". https://www.mdpi.com/2078-1547/8/2/32
2. Harper, C.D.; Hentrickson, C.T.; Mangones, S.; Samaras, C. Estimating potential increases in travel with autonomous vehicles for the non-driving, elderly and people with travel-restrictive medical conditions. Transp. Res. Part C 2016, 72, 1–9.
3. Yap, M.D.; Correia, G.; Van Arem, B. Valuation of travel attributes for using automated vehicles as egress transport of multimodal train trips. Transp. Res. 2015, 10, 462–471.
4. V. V. Dixit, S. Chand, and D. J. Nair, "Autonomous vehicles: disengagements, accidents and reaction times," PLoS one, vol. 11, no. 12, p. e0168054, 2016
5. Chen, R.B.; Armington, W. Household activities and travel patterns with autonomous vehicles: In-vehicle activity decisions. In Proceedings of the TRB 95th Annual Meeting, Washington, DC, USA, 10–14 January 2016.
6. Posenethttps://github.com/kentsommer/tensorflow-posenet
7. Posenet-for-installationshttps://github.com/oveddan/posenet-for-installations
8. Open pose https://github.com/CMU-Perceptual-Computing-Lab/openpose
9. Chainer_realtime_multi-Person_Pose_Estimation-masterhttps://github.com/DeNA/Chainer_Realtime_Multi-Person_Pose_Estimation
10. Vogelpohl, Matthias Kühn, Thomas Hummel, Tina Gehlert, Mark Vollrath (2018). Transitioning to manual driving requires additional time after automation deactivation., Transportation Research Part F:55 (2018),464-482. https://www.sciencedirect.com/science/article/pii/S1369847817301924
11. Merat, N., Jamson, A. H., Lai, F. C., Daly, M., & Carsten, O. M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. Transportation Research Part F: Traffic Psychology and Behaviour, 27, 274–282. https://www.sciencedirect.com/science/article/pii/S1369847814001284
12. Gold C, Dambock D, Lorenz L, et al. "Take over!" How long does it take to get the driver back into the loop[J]. Proceedings of the Human Factors & Ergonomics Society Annual Meeting, 2013, 57(1):1938-1942.
13. Lex Frdman (2017), MIT Autonomous Vehicle Technology Study:Large-scale Deep Learning Based Analysis of Driver Behaviour and Interaction with Automation. arXiv:1711.06976v1.
14. Castro (2015), Predicting Daily Activities From Egocentric Images Using Deep Learning. Proc Int Symp Wearable Comput. 2015 Aug; 2015: 75–82.

15. Braunagel (2015), Driver-Activity Recognition in the Context of Conditionally Autonomous Driving. ITSC 2015
16. Wang (2013), An approach to pose-based action recognition. CVPR 2013
17. Du (2015), Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. CVPR 2015