RECOGNITION OF MANUAL DRIVING DISTRACTION THROUGH DEEP-LEARNING AND WEARABLE SENSING

Li Li[a], Ziyang Xie[b], Xu Xu[a]
[a]Edward P. Fitts Department of Industrial & Systems Engineering
[b]Mechanical and Aerospace Engineering
North Carolina State University, Raleigh, NC, USA
Email: xxu@ncsu.edu

Summary: The goal of this study is to design a novel framework incorporating deep-learning techniques and wearable sensors to recognize manual distractions during driving. Manual distraction is defined as hands off the wheel for any reason (e.g. trying to get a cell phone). In this preliminary study, participants were tasked to drive in city street and highway scenarios in a driving simulator. Verbal instructions prompted participants to perform various manual distraction tasks. The motion of driver's right wrist during driving was recorded by a wearable inertial measurement unit. A deep-learning technique called convolutional neural network (CNN) was then constructed and trained based on 72% of the experiment trials, and evaluated by the remaining 28% of trials. The results indicated that the convolutional neural network is able to recognize the type of manual distraction task based on the right wrist motion with 87.0% accuracy and F1-score of 0.87. The results indicated that there is a good potential to apply deep-learning techniques and wearable sensing to monitor driver's inattention status.

INTRODUCTION

Driving distraction is a critical issue in transportation safety. In 2012, distracted driving was associated with 3,328 deaths and 421,000 injuries in the U.S. (NHTSA, 2014). Adequate distraction mitigation strategies have been shown to improve driving performance (Donmez, Boyle, & Lee, 2006). Moreover, in recent years, automated driving systems have received a great deal of attention. Driving distraction status has been considered as a key factor for assessing the timing of transfer of control between manual and automated driving (Nilsson, Falcone, & Vinter, 2015). Thus, there is an urgent need to provide an effective method for understanding drivers' real-time attentional status in order to improve transportation safety.

Driving distraction is defined as a diversion of attention away from activities for safe driving toward competing activities (Lee, Regan, & Young, 2008). Specifically, distractions can be categorized into three types: 1) manual distractions, such as hands off the wheel; 2) visual distractions, such as eyes off the road, and 3) cognitive distractions, such as recalling someone's telephone number. Manual distraction is commonly combined with other types of distractions, as manual distraction is a triggered response of other types of distraction (Regan, Hallett, & Gordon, 2011). Therefore, recognition of manual distraction can contribute to the monitoring of overall distracted driving.

Human kinematics variables used for manual distraction recognition generally include one or more measures of position and angular orientation of body segments and joints. An intuitive

method to recognize manual distractions is to use the location of hands relative to other body segments (Gallahan et al., 2013). Taking this approach a step further, jointly using hand and elbow kinematics can differentiate, for example, cell phone use from combing hair.

While there is great potential for using a driver's in-car movements to infer manual distractions, the spatial features and linear classifiers only provide limited recognition accuracy and robustness. Since manual distractions generate unique temporal patterns of human kinematics, the time domain also includes substantial information for manual distraction recognition. The combination of spatial and temporal features of human kinematics have been previously implemented in the identification of human activities (Bourke, O'donovan, & Olaighin, 2008). Very recently, deep-learning techniques, such as convolutional neural network (CNN) have shown a great potential in recognizing human daily activities (Ordóñez & Roggen, 2016).

In terms of driver motion tracking, there have been many studies making use of other devices, such as RGB / RGB-Depth sensors (Kurakin, Zhang, & Liu, 2012), or wearable inertial measurement units (IMUs) (Altun, Barshan, & Tunçel, 2010), to track human kinematics and reconstruct human pose. Among these devices, wearable IMUs have some unique advantages for measuring in-car human movement. First, unlike fitting camcorders or depth sensors in a vehicle, using IMUs does not require any retrofitting of a car cab. Second, wearable IMUs can track driver in-car movement for persons frequently driving multiple cars. Third, wearable IMUs are now becoming commonplace, with advent and popularization of fitness and activity trackers.

In this study, drivers drove through a number of different simulated driving environments while performing a variety of in-vehicle, secondary activities. A CNN was then constructed and trained by 72% of the driver motion data (from 16 subjects) collected through a wearable IMU. The performance of this deep neural network on recognizing distracted behavior during driving was then evaluated by remaining 28% of the dataset (4 subjects).

METHOD

Participants

The kinematics data of 20 participants (8 females and 12 males, age range between 25 to 51 years old) during driving were adopted in this preliminary analysis. All the participants have a valid drivers' license, normal or corrected to normal visual acuity, and no disposition to motion sickness.

Apparatus

The study was performed in an RTI driving simulator (Realtime Technologies, Ann Arbor, MI), which is a fixed-base simulator that consists of an open-cab vehicle mock up, including accelerator and brake pedals, steering wheel, dashboard, instrument panel, and center console. The driving environments were presented on three 46-inch widescreen LCD displays. (Figure 1). Various driving environments and traffic scenarios were generated using RTI SimCreator and SimVista software. The movements of driver's trunk and wrist were measured by two wearable IMUs (MVN, Xsens technologies, the Netherlands).

Figure 1. **Experiment Setup. The figure shows the participant is using the navigation panel while driving on highway scenario with right wrist kinematics tracked by an IMU sensor**

Driving scene

In order to collect a more representative dataset, a variety of simulated driving scenarios were used for the study, including city streets and highway scenarios. In the city-street scenario, the participants drove on straight/curved road through an urbanized setting, including traffic signals, ambient traffic and a mixture of residential and commercial buildings. Participants were required to make right/left turns at intersections, based on directional arrows that were presented on the forward road. . In the highway scenario, the participants drove on a three-lane divided highway, including a mixture of straight and curved roads and with other ambient traffic.

Distraction tasks

Five types of manual distraction tasks were selected for this study: During the driving experiment trials, five types of manual distraction tasks were randomly assigned to the participants. The five designed tasks are cell phone talking (*Phone*), cell phone texting (*Text*), drinking water (*Drink*), using navigation panel (*TouchScreen*), and placing a marker pen into the cup holder (*Marker*).

Experiment protocol

At the start of the session, participants completed an informed consent and demographic questionnaires. The inertial sensors were then attached on the participants' right wrist and trunk. The reason for recording the motion of right wrist is because all the distraction tasks assigned to subjects involve use of right hand. Before conducting the experimental blocks, participants completed a practice trail in order to familiarize themselves with and acclimate to the simulator. During the experimental blocks, drivers randomly performed the different distraction tasks, based on auditory instructions. Tasks were self-paced and, once the assigned task was complete, the participant pushed a button on the steering wheel. Another verbal instruction was presented approximately 20 seconds later. Drivers complete six experimental blocks, each of which lasted for approximately 20 minutes. One of these six blocks did not include any distraction task and

were used as the baseline to enhance the performance of classification. In between each block, drivers were given a five-minute break.

Data Processing

The right wrist XYZ positions and angular orientation described by quaternion were extracted from the raw data of wearable IMU sensors. Specifically, the right wrist kinematics was described with respect to the trunk coordinate system, which is defined by International Society of Biomechanics standard. (Wu et al., 2005). The kinematics data during distraction tasks were manually segmented into three segmentations by an experimenter based on the following four key events: Start, Initiate, Return, and End. Start represents the frame that the participant starts moving their hand off from the steering wheel and toward the target of the assigned distraction task (e.g., starting to reach towards touchscreen). Initiate represents the frame that the participants make initial contact with the target of assigned distraction task (e.g., first touch of the touchscreen). Return represents the frame that the participant finishes the task and starts to return from the target. End represents the frame that the participants' hand moves back to the initial position on the steering wheel. Considering the great inter-participants variability regarding how each distraction task was performed, we chose the segments from Start to Initiate as the target pattern to be classified, which were more consistent from one participant to another. In addition, this segment represents the beginning of the distraction task while driving, so recognizing this segment will better facilitate real-time activity recognition of distracted driving in the future work. Figure 2 shows sample spatiotemporal patterns from Start to Initiate for the assigned two distraction tasks. From all the twenty participants, 395 *Phone* events, 378 *Text* events, 396 *Drink* events, 387 *TouchScreen* events, and 393 *Marker* events were extracted in total from the continuous kinematics data during driving. In addition, 393 kinematics segments during normal driving were randomly extracted to form a *DrivingOnly* category. All 2342 data segments are normalized to the same length so that they can be imported to a CNN.

Convolutional Neural Network

The convolutional neural network (CNN) is a type of deep neural network. Previous studies have shown that the CNN is good at pattern recognition and classification for high dimensional data. Given human kinematics data are typically multidimensional, CNN would be a good candidate for human activity recognitions (Yang, Nguyen, San, Li, & Krishnaswamy, 2015). In this study, the first few layers of CNN convolve the input data with convolutional kernels to extract low-level features. For signals collected from the inertial sensor, the features are the spatial relations between output from different channels (viz. 3 channels for linear position, and 4 for angular orientation as quaternion). The convolutional layers are connected to fully connected layers (FC), which reshape the data into a one-hot vector indicating the inferred body motion. We adopted AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) as our base model, which was originally designed for recognition in image classification, and consists of 5 convolutional layers and 2 fully connected layers. The first layer of the original AlexNet takes 3-channel RGB images as input, and convolve it with (11, 11, 3) kernels. However, our data is 7-channel time series, so the receptive field is smaller. Therefore, we changed the kernel size of convolutional layer one and layer two to (3, 3) and (7, 7) respectively. Additionally, a ReLU activation function was used at each convolutional layer to increase nonlinearities, and a max pooling layer was added behind the first, the second, and the fifth layer for down-sampling the data.

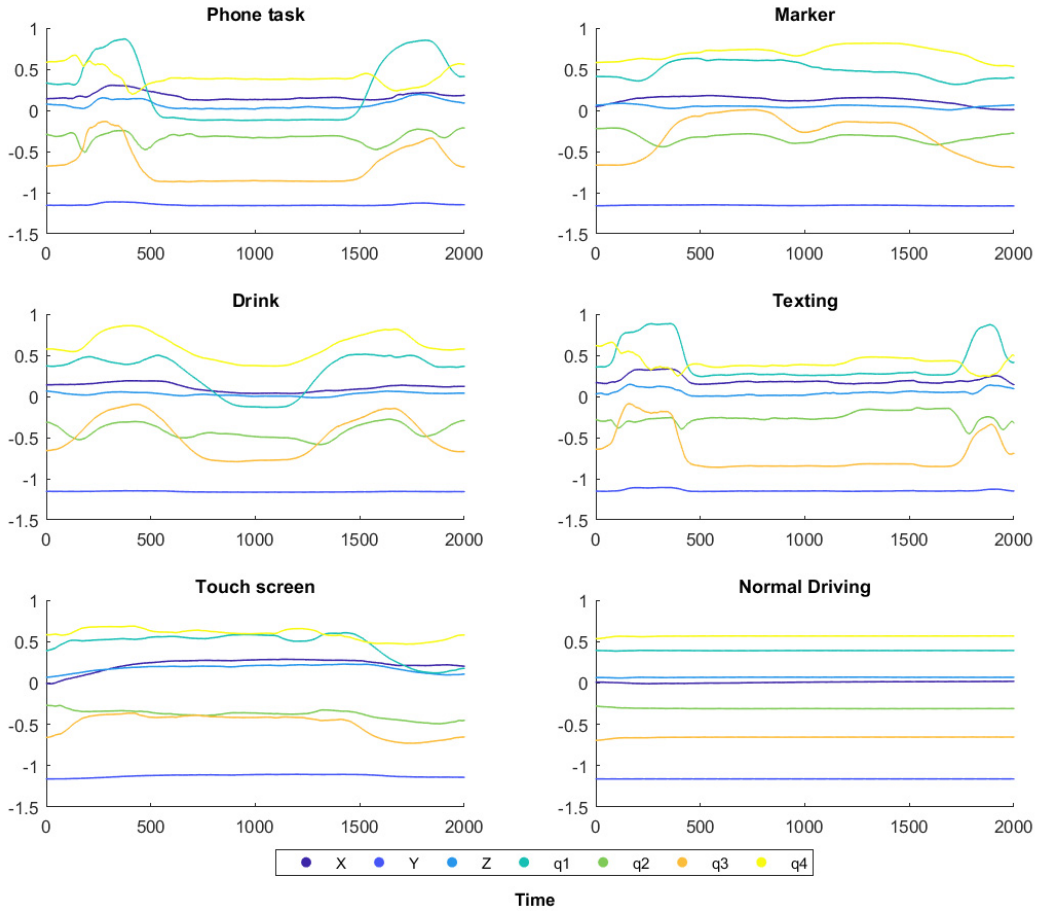

Figure 2. **An example (Subject 01 in training set) of right wrist kinematics during different tasks. The time span of each task was normalized to 0-2001 using linear interpolation.**

Among 2342 data segments, 1688 segments from 16 randomly selected participants were used for CNN training and 654 segments from the rest 4 participants were used for testing. The whole network was trained on 2 Titan V GPU cards. Learning rate and batch size were set as hyper-parameters. The cross entropy of softmax was adopted as the loss function. AdamGrad was used as the optimizer.

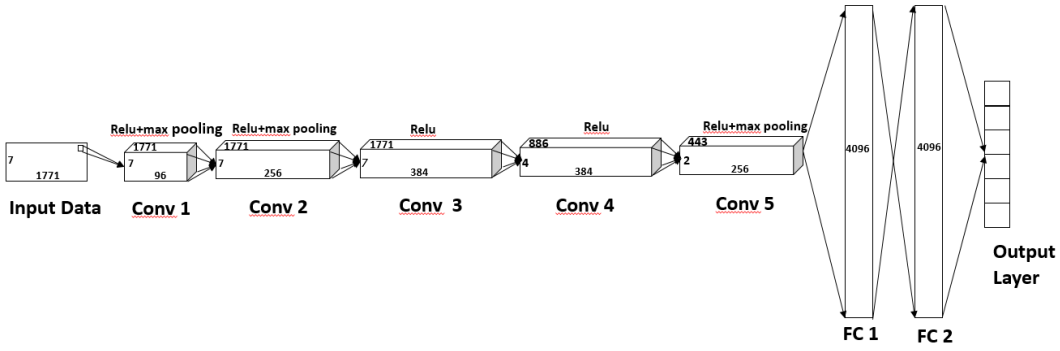

Figure 3. Architecture of modified AlexNet

RESULTS

The recognition outcome from the testing dataset is shown in Table 1. The overall accuracy is 87.0 % and the F1-score, which considers both the precision and recall of the test, is 0.87. The formulas for calculating recall, precision, and F1-score are given in the following:

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad (1)$$

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (2)$$

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (3)$$

Table 1: Confusion Matrix of the classification results using CNN. Rows represent the actual class and columns represent the predicted class. The diagonal entries show the number of samples correctly classified.

	Phone	*Text*	*Drink*	*TouchScreen*	*Marker*	*DriveOnly*	Recall
Phone	80	0	11	20	0	1	72%
Text	2	97	1	0	9	0	89%
Drink	0	0	102	2	5	0	94%
TouchScreen	0	0	5	94	1	7	88%
Marker	1	0	5	1	101	1	93%
DriveOnly	2	0	6	0	6	95	87%
Precision	94%	100%	78%	80%	83%	92%	

DISCUSSION

The preliminary results from this study indicate that CNN and wearable sensors may provide a robust method to infer manual distractions during driving, given the high accuracy and F1-score in driver body movement classification.

However, it should be noted that these results need to be interpreted with caution. First, only a limited number of participants have been recruited. Among the overall population, the body movement during distracted driving could be very different from person to person. Thus, the variability in driver's body motion may not be well presented in the training dataset. Additional data from more drivers are needed for training a more robust CNN. Second, in this preliminary study the time window length for each distracted body motion is the same. Adjusting the time window for improving the classification results is underway now. Third, AlexNet was adopted in this study for a quick result delivery. For one thing, considering its good performance in large-scale image classification task, the structure of it could be simplified while keeping the same performance on body motion classification since the input data has less features and variability, and the some kernels are possibly becoming redundant. For another, the spatial relations along the column side of the data is dependent on how we stack the data from seven channels so that it is possible to have structures other than CNN that work better in body motion classification. In the future work, more advanced structure of neural network will be adopted to keep improving the robustness of the method.

CONCLUSION

This paper has proposed a novel framework incorporating deep-learning techniques and wearable sensors to recognize manual distractions during driving. More specifically, the CNN (AlexNet) was improved and fine-tuned for this task. The results indicate that the CNN is able to detect and classify different type of manual distractions with satisfying sensitivity. Future work will investigate the possibility of using more advanced and simpler network structure, and address the issue of dynamic window length for different motions.

ACKNOWLEDGMENTS

This manuscript is based upon work supported by the National Science Foundation under Grant No 1822477.

REFERENCES

Altun, K., Barshan, B., & Tunçel, O. (2010). Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, *43*(10), 3605–3620.

Bourke, A. K., O'donovan, K. J., & Olaighin, G. (2008). The identification of vertical velocity profiles using an inertial sensor to investigate pre-impact detection of falls. *Medical Engineering & Physics*, *30*(7), 937–946.

Donmez, B., Boyle, L. N., & Lee, J. D. (2006). The impact of distraction mitigation strategies on driving performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(4), 785–804.

Gallahan, S. L., Golzar, G. F., Jain, A. P., Samay, A. E., Trerotola, T. J., Weisskopf, J. G., … Ieee. (2013). *Detecting and Mitigating Driver Distraction with Motion Capture Technology: Distracted Driving Warning System*. *2013 Ieee Systems and Information Engineering Design Symposium*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kurakin, A., Zhang, Z., & Liu, Z. (2012). A real time system for dynamic hand gesture recognition with a depth sensor. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European* (pp. 1975–1979). IEEE.

Lee, J. D., Regan, M. A., & Young, K. L. (2008). Defining driver distraction. In M. A. Regan, J. D. Lee, & K. L. Young (Eds.), *Driver distraction: theory, effects, and mitigation*. Boca Raton, FL: CRC Press.

NHTSA. (2014). Distracted driving 2012. Washington, DC.

Nilsson, J., Falcone, P., & Vinter, J. (2015). Safe transitions from automated to manual driving using driver controllability estimation. *IEEE Transactions on Intelligent Transportation Systems*, *16*(4), 1806–1816.

Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, *16*(1), 115.

Regan, M. A., Hallett, C., & Gordon, C. P. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis & Prevention*, *43*(5), 1771–1781.

Wu, G., van der Helm, F. C. T., Veeger, H. E. J., Makhsous, M., Van Roy, P., Anglin, C., … Buchholz, B. (2005). ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion - Part II: shoulder, elbow, wrist and hand. *Journal of Biomechanics*, *38*, 981–992.

Yang, J., Nguyen, M. N., San, P. P., Li, X., & Krishnaswamy, S. (2015). Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *Ijcai* (Vol. 15, pp. 3995–4001).