WatchUDrive: Differentiating Drivers and Passengers Using Smartwatches

Alex Mariakakis
Computer Science and Engineering
University of Washington
Seattle, Washington 98195
atm15@cs.washington.edu

Vijay Srinivasan, Kiran Rachuri,
Abhishek Mukherji
Samsung Research America
Mountain View, CA 94043
{v.srinivasan, k.rachuri, a.mukherji}@samsung.com

***Abstract*—Personalization and automation in future smart vehicles hinge on accurately identifying the driver and passengers in the vehicle. Traditional approaches either require additional infrastructure or impose assumptions about how users interact with their smartphones. The recent proliferation of commercial smartwatches enables new opportunities to solve this problem due to the fixed position of the watch on the wrist. We use this observation to motivate WatchUDrive, our smartwatch-based application for identifying whether the wearer is the driver or a passenger in a vehicle. We evaluate two smartwatch sensing modalities for driver vs. passenger differentiation: the accelerometer and the camera. Using 40 in-vehicle episodes collected from 8 users and 8 different vehicles, we show that the accelerometer yields 90% accuracy within 10 seconds, whereas the camera only yields 62% accuracy within 110 seconds.**

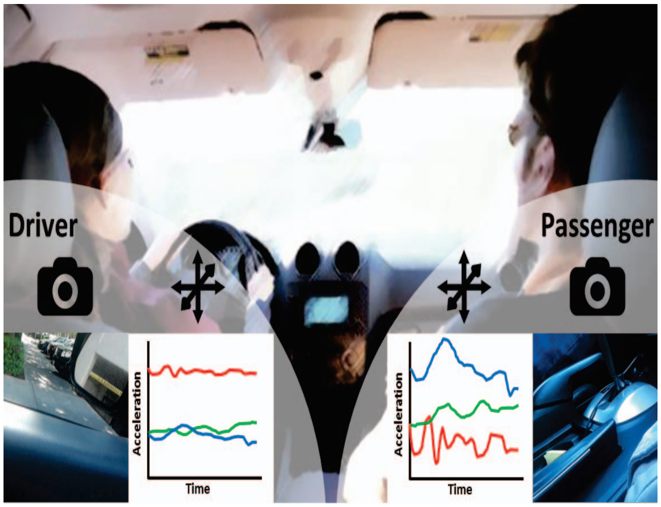

Fig. 1. WatchUDrive identifies the user as either the driver or a passenger in the vehicle using a smartwatch. The accelerometer senses the orientation of the user's arm, while the camera observes the user's immediate surroundings.

I. Introduction

Distracted driving has become a great area of concern since the widespread adoption of mobile phones. The National Safety Council estimates that 1 in every 4 vehicle crashes involves some sort of mobile phone use [6]. Companies have created Bluetooth-based solutions for phones to automatically silence themselves when the user is in a vehicle [1], but silencing the phone for all users including passengers is inconvenient. To solve this problem and support other smart vehicle personalization features (*e.g.*, seat position presets), we need approaches to automatically identify if a user is a driver or a passenger. Existing approaches for driver-passenger differentiation suffer from several drawbacks; some approaches require additional hardware infrastructure or calibration per vehicle [3], [8], other approaches makes unrealistic assumptions about how the phone is always carried by users [10], [4], while yet another group of approaches require data from multiple sources in the car [5], [9].

In this paper, we propose WatchUDrive (Figure 1), a system that uses only the sensors available on a smartwatch for driver-passenger differentiation. By using a smartwatch, we are able to overcome many of the drawbacks underlying prior work while maintaining comparable accuracy. We explore the use of two sensing modalities in this work to differentiate the driver and passenger. First, we show how the accelerometer can be used to sense the orientation of the user's arm. Second, we explore how the camera can capture the user's immediate surroundings. We apply a dynamic prediction aggregation scheme that combines individual predictions while maintaining a balance between accuracy and latency. After all, neither an incorrect prediction nor a correct prediction made ten minutes into an episode are useful for a system that is meant to start at the beginning of a driving episode. We evaluate WatchUDrive using a dataset of 40 in-vehicle episodes collected from 8 users and 8 different vehicles. Through our analysis, we show that using just the accelerometer achieves 90% accuracy in as little as 10 seconds, whereas the camera only achieves 62% accuracy within 110 seconds. We conclude briefly with our observations for why this discrepancy occurs and other considerations.

II. Related Work

Prior work in driver-passenger differentiation generally utilize motion sensors, audio sensors, or some combination of the two.

He *et al.* [5] observe that users in the same vehicle experience different magnitudes of acceleration during particular driving events. When going over a speedbump or a pothole, a person sitting in the front of the vehicle will experience the bump before a person sitting in the back; during a turn, a

978-1-5090-1941-0/16/$31.00 ©2016 IEEE

person on the inner radius of the turn will experience less centripetal acceleration than a person sitting on the outer radius. Wang *et al.* [9] replace the need of a collocated user with an OBD-II port adapter. These works require data to be shared from multiple sources and can only identify the side in which the user is seated (*i.e.*, right or left, front or back) unless certain conditions are met.

Chu *et al.* [2] listen for inertial microevents that are unique to vehicles. For example, they observe that the pressing of the gas pedal can be detected with the smartphone's accelerometer if the phone is in the right pocket. Their work also relies on the detection of audio microevents, like the clicking of the turn signal, to localize the user. The inertial microevents are dependent on the occurrence of these infrequent microevents to achieve high accuracy and require knowledge about the position of the phone on the user's body.

Yang *et al.* [10] and Gruteser *et al.* [4] produce custom artificial inaudible tones from different channels of the vehicle's stereo station and use time-of-flight comparisons for localization. Feld *et al.* [3] provide a more hardware-centric solution. They install a high-fidelity directional microphone into each seat. By training a voice profile for each person who may enter the vehicle, a simple one-to-one mapping can be generated between voice and microphone. Swerdlow *et al.* [8] consolidate the hardware setup to microphone arrays at the front and back of the vehicle; localization is performed by measuring the time-delay-of-arrival between different microphones within the same array. Solutions in this category require either additional hardware or coordination between the user's phone and the vehicle.

WatchUDrive is free from many of these limitations. The entire system can run on a smartwatch alone, although pairing with a smartphone would extend the watch's battery life. It also does not rely on microevents that occur every few minutes and can thus produce predictions much quicker. Finally, the fact that the smartwatch is worn on the user's wrist eliminates any the need for assumptions about how the user handles the device.

III. Design

In this section, we outline the two sensing approaches we explored for driver-passenger differentiation using the accelerometer and camera respectively. We then discuss the dynamic scheme we use to combine instantaneous predictions into a single, episode-level prediction.

A. Accelerometer Classifier

One of the most obvious differences between drivers and passengers is the act of driving itself. As a driver steers the vehicle, they must keep hold of the steering wheel for long periods of time. Passengers, on the other hand, are less restricted in how they keep their hands.

We measure the wrist's orientation by recording the effect of gravity on the smartwatch's accelerometer. A common way of expressing an object's orientation is through the Euler angles: roll (α), pitch (β), and yaw (γ). Calculating the yaw, or the rotation around the object's z-axis, requires a magnetometer in order to ground the rotation to some reference. Commodity smartwatches currently do not have magnetometers, so we are limited to measuring the wrist's roll and pitch. After passing the accelerometer measurements through a low-pass filter, we calculate the median roll and pitch of the user's wrist over a sliding five second window with 50% overlap (parameters which were determined through experimentation). We discard segments where the acceleration variation is large since the median is not a proper representation of those segments.

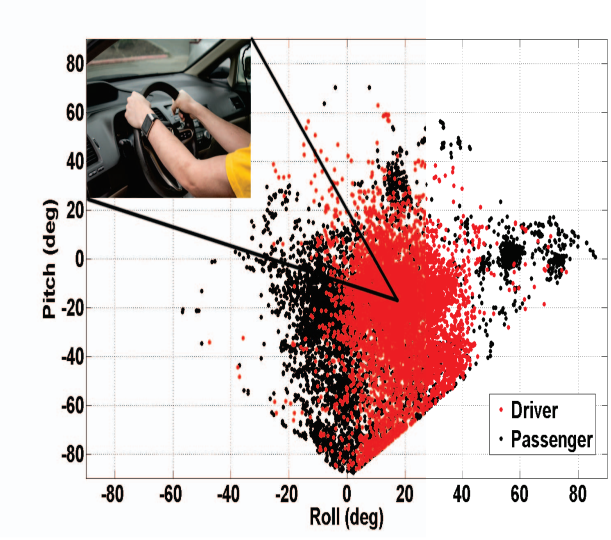

Fig. 2. The distribution of wrist orientation vectors for drivers and passengers across our dataset.

Figure 2 shows the how the orientation of the user's wrist varies between drivers and passengers from our dataset. The dense red region marked for drivers highlights the "9-and-3" manner of gripping a steering wheel as taught by driving instructors. In this posture, the driver's hands are higher than their elbows, creating a tilt that is hypothetically unique to driving. For passengers, we can see that the distribution of orientation vectors is much more scattered. This is to be expected since passengers have much more freedom with their hands as the vehicle is moving.

B. Camera Classifier

To handle the cases when users hold the steering wheel in a different manner, we explored the capabilities of the camera. The cameras available on smartwatches tend to point away from the outside edge of the user's wrist. The camera on the driver's watch is often elevated enough so that it captures a view of the outdoors, the driver-side mirror, or the middle of the door frame. Passengers' wrists are often not elevated enough to get such an interesting view. If it does happen, though, passengers on the right side often view the interior of the vehicle. One possible way to distinguish between images from both cases would be to identify a few objects that are common across most vehicles (*e.g.*, gear stick, side mirrors); however, identifying these objects would require vehicle-specific training since every manufacturer has its own distinct style. We avoid these issues by relating the classification of these images, or vehicle scenes, to the problem of outdoor scene recognition. Instead of creating individual classifiers for trees, buildings, *etc.*, Oliva and Torralba proposed a low dimensional representation of a scene's spatial structure, leading to what are now known as GIST features [7]. At a high level, GIST features capture the dominant spatial structure of a scene using coarsely localized gradient histograms. Since GIST features are roughly dependent on the orientation of the image, we align images in the same frame of reference using the accelerometer vector recorded when the image was captured. Oliva and Torralba demonstrate how GIST features can capture characteristics like openness, roughness, and expansion, a vocabulary which can also be applied to vehicular scenes. On the left side of Figure 1, the driver image has a high degree of openness since there is a distinct division formed between the vehicle door and the window. The passenger image on the right side of Figure 4 1, on the other hand, has a high degree of roughness due to the complexity of the vehicles controls.

The camera often has an uninformative view of the user's surroundings when it is direct downward or in the user's lap; in these cases, either the user's clothing is captured or the camera is completely obstructed. To quantify whether a view is informative or not, we first use localized grayscale and color histograms as features. We then filter out images with little variance across the GIST, grayscale, and color features of each cell in the 4×4 grid used to create them, leaving images that typically contain some distinct aspect, such as the edge of a seat or a close-up view of a door. We also remove any images captured as the user is moving their arm, which tend to be blurry and lead to poor classification.

C. Prediction Aggregation

The accelerometer and camera classifiers each produce an individual, independent classification prediction for the user's state. We assume that people remain in the same seat over the course of an episode (the time between the user entering and leaving a vehicle), allowing us to combine individual predictions into a more robust episode-level prediction.

We define $p_{\mathrm{D}}(t)$ to be the probability that the user is a driver at instance t. This probability can be computed independently by a classifier every t_s seconds, which we will call the prediction sampling period. We define a range of time, t_{min} to t_{max}, to be the minimum and maximum time that a classifier can take before reaching a decision. We then aggregate individual predictions into a single value $\overline{p}_{\mathrm{D}}$ by taking the mean of all the individual predictions $p_{\mathrm{D}}(t)$ up until some time t where $t_{min} < t < t_{max}$. At every moment t, we inspect the value $\overline{p}_{\mathrm{D}}$ and assign a label to the user if its value exceeds a threshold probability Φ. The user is labeled as a driver if the overall probability satisfies $\overline{p}_{\mathrm{D}} \geq \Phi$ or as a passenger if it satisfies $\overline{p}_{\mathrm{D}} \leq 1 - \Phi$. If either condition is met, no more samples are collected from the watch and we proactively turn off WatchUDrive. This is done to conserve the smartwatch's battery power. If a prediction is not made by t_{max}, we accept the prediction according to the $\overline{p}_{\mathrm{D}}$ regardless of its magnitude relative to random chance. Together, t_s, t_{min}, t_{max}, and Φ are the parameters that can be adjusted to vary the accuracy, precision, recall, and latency of the system.

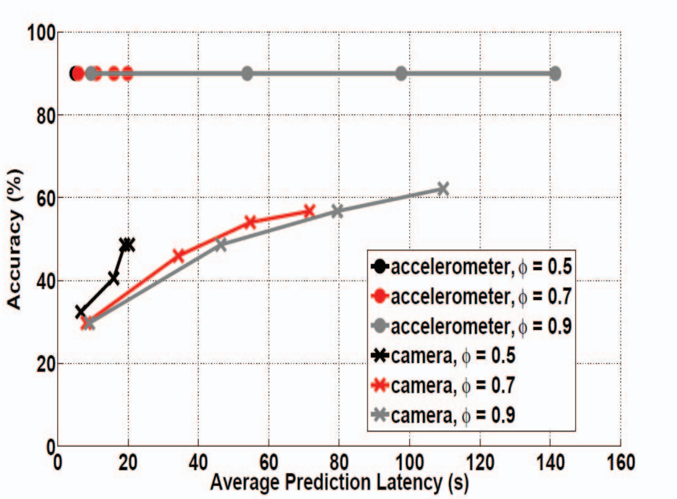

Fig. 3. The effect of Φ on the prediction latency and accuracy associated with the accelerometer and camera classifiers.

IV. Evaluation

A. Data Collection

We gathered 40 episodes for evaluation using Samsung Gear 2 devices paired with a Samsung Galaxy S3 or S4 smartphone for the sake of large data file storage. Out of the 40 episodes, 20 were from drivers and 20 were from passengers. Within the latter group, 10 sat in the front passenger seat, 6 sat in the back right seat, and 4 sat in the back left seat. We believe this reflects the general distribution of seating behavior in vehicles. Episodes were collected from 8 different users, 8 different vehicles, and span both day and night driving. Our primary evaluation metric is episode-level classification accuracy through leave-one-out episode-level classification. One at a time, each episode is removed, leaving the remaining episodes to train models for both the wrist orientation and the seat view. We use Naïve Bayes for our classifiers since it provides probabilistic outputs by default.

B. Results

Figure 3 is a parametric graph that shows the accuracy versus latency tradeoff of the accelerometer and camera-based classifiers. For these lines, t_s = 5 seconds and t_{min} = 5 seconds. Both the accuracy and the average prediction latency are dependent variables effected by the independent variable t_{max}, which varies between 10-180 seconds.

We found that our accelerometer classifier achieved 90% accuracy for all values of t_s, t_{min}, t_{max}, and Φ that we picked, which can be attributed to the high performance of the classifier itself rather than the dynamic prediction aggregation scheme. Using our leave-one-out training approach, we had sufficient data to identify natural postures for both drivers and passengers. The dynamic prediction scheme does, however, have an effect on how long the system takes before a classification is finalized. As Φ increases, it is harder for the aggregate probability to satisfy the corresponding threshold. This means that some episodes take longer to converge to a final prediction, thereby increasing the average prediction latency. The accuracy remains at 90% despite this trend, so selecting small values t_s, t_{min}, t_{max}, and Φ for the accelerometer-based classifier is optimal for low latency predictions.

The parameters of the prediction aggregation scheme have a more pronounced effect on the accuracy of the camera classifier. We observe that the overall accuracy is worse than random chance early on in the episode, but steadily improves as time progresses; this could be due to the atypical behaviors and scenery at the beginning of the episodes (*e.g.*, parking decks, lots of turning). Just as with the accelerometer classifier, we observe that the latency increases as Φ increases since it is harder for episode predictions to satisfy that constraint. However, unlike the accelerometer classifier, we observe that the accuracy also increases with Φ as higher confidence predictions are obtained after the initial noise period. We achieve the best results of 62% accuracy with our camera classifier with an average prediction latency of roughly 110 seconds given that Φ = 0.9 and t_{max} = 180 seconds. Although further increasing t_{max} improves accuracy even more, doing so reduces the efficacy of our system for applications.

V. Discussion

In the end, we were unable to use the camera sensor to improve on the high accuracy achieved by the accelerometer. We tried to combine the predictions both of the sensors to a single, unified classifier, yet we were unable to identify a weighting such that the overall accuracy improved over the accelerometer alone. After inspecting the data, we found that the moments when the accelerometer classifier failed were just a subset of the moments when the camera classifier failed as well. One such situation is when the driver grips the wheel with only the hand that is not wearing the watch. The arm with the watch in this scenario usually rests in the driver's lap, which according to the accelerometer looks similar to how passengers keep their arms. We had hoped that the camera would prove beneficial here by distinguishing between the driver's and passenger's surroundings, but we found that the camera instead often captured parts of the user's body.

A large factor in the performance difference between the two classifiers is the variability in the features used. When considering the intuition behind the accelerometer classifier, there are only so many natural positions for a driver to hold their arm. The camera classifier, on the other hand, is susceptible to far more variability in the data. The user's clothing, the vehicle's interior, and even the time of day all contributed to the images captured by the camera. One way to remedy this issue would be to consider a user-specific system, which would reduce some of that variability. Power consumption is another dimension where the accelerometer classifier is superior to the camera classifier. This can be attributed to the lower power consumption of the sensors themselves and the shorter prediction latency.

The placement of the wrist-worn camera was a make-or-break factor in the success of the camera classifier. The camera in its current placement is intended for intentional image capturing, not passive, continuously running applications. We hypothesized that the placement would have been beneficial for sensing an activity like driving where there is an expected behavior of the user's arms. In many cases, however, the camera was either too close to the side of the vehicle, covered by an article of clothing, or pointing directly downward. A camera that points out of the smartwatch's face would likely be obstructed less often; however, this removes the ability of the user to see the camera's view as they are actively taking a picture.

VI. Conclusion and Future Work

In this paper, we presented WatchUDrive, a smartwatch system for differentiating vehicle drivers and passengers. We examined two sensor modalities to solve this problem: the acceleromter and the camera. Our accelerometer classifier achieved 90% accuracy within 10 seconds, which we believe is strong enough to merit deployment; our camera-based classifier achieved lower accuracy and could not be used to improve upon using the accelerometer alone. Despite these results, we suspect that the camera will be useful in other activity recognition scenarios. One scenario we are currently exploring is food journaling; a wrist-worn camera could capture the plate of food in front of the user depending on the orientation of the wrist. We plan to further explore this scenario and others in future work on smartwatch-based context recognition.

References

[1] Automatic Labs. Automatic: An Auto Accessory to Make You a Smarter Driver, 2015.

[2] Hon Lung Chu, Vijay Raman, Jeffrey Shen, Romit Roy Choudhury, Aman Kansal, and Victor Bahl. Poster: you driving? talk to you later. In *MobiSys*, pages 397–398, 2011.

[3] M Feld, T Schwartz, and C Müller. This is me: using ambient voice patterns for in-car positioning. In *Ambient Intelligence*, pages 290–294, 2010.

[4] Marco Gruteser, Richard Paul Martin, Chen YingYing, and Jie Yang. Systems and methods for detecting driver phone use leveraging car speakers, 2013.

[5] Z He, J Cao, X Liu, and S Tang. Who sits where? Infrastructure-free in-vehicle cooperative positioning via smartphones. In *Sensors*, pages 11605–11628, 2014.

[6] National Safety Council. Distracted Driving: One Call Can Change Everything, 2015.

[7] A Oliva and A Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[8] A Swerdlow, K Kroschel, and T Machmer. Speaker localization in vehicles via acoustic analysis. In *DAGA*, 2007.

[9] Y Wang, J Yang, H Liu, and Y Chen. Sensing vehicle dynamics for determining driver phone use. In *Proc. MobiSys '13*, pages 41–54, 2013.

[10] Jie Yang, Simon Sidhom, Gayathri Chandrasekaran, Tam Vu, Hongbo Liu, Nicolae Cecan, Yingying Chen, Marco Gruteser, and Richard P Martin. Detecting driver phone use leveraging car speakers. In *Proc. MobiCom '11*, pages 97–108, 2011.