Predicting Perceived Visual and Cognitive Distractions of Drivers with Multimodal Features

Nanxiang Li, *Student Member, IEEE,* Carlos Busso, *Senior Member, IEEE,*

***Abstract*—The driver's behaviors can be affected by visual, cognitive, auditory and manual distractions. While it is important to identify the patterns associated with particular secondary tasks, it is more general and useful to define distraction modes that capture the general behaviors induced by various sources of distractions. By explicitly model the distinction between types of distractions, we can assess the detrimental effects induced by new in-vehicle technology. This study investigates drivers' behaviors associated with visual and cognitive distractions, both separately and jointly. External observers assessed the perceived cognitive and visual distractions from real world driving recordings, showing high inter-evaluator agreement in both dimensions. The scores from the perceptual evaluation are used to define regression models with elastic net regularization and binary classifiers to separately estimate the cognitive and visual distraction levels. The analysis reveals multimodal features that are discriminative of cognitive and visual distractions. Furthermore, the study proposes a novel joint visual-cognitive distraction space to characterize driver behaviors. A data-driven clustering approach identifies four distraction modes that provide insights to better understand the deviation in driving behaviors induced by secondary tasks. Binary and multi-class recognition problems demonstrates the effectiveness of the proposed multimodal features to infer these distraction modes defined in the visual-cognitive space.**

***Index Terms*—Driver cognitive distraction, driver visual distraction, human evaluation, stepwise regression, logistic regression, binary classification.**

I. Introduction

THe U.S. Department of Transportation published a report with voluntary guidelines for the automobile industry to design technologies that are safe for drivers [1]. These guidelines respond to the concerns about the detrimental effects on driving behaviors induced by in-vehicle technology and mobile devices. Despite the effort, the National Highway Traffic Safety Administration (NHTSA) estimates that the number of traffic deaths has increased 5.3 percent in 2012. Driver's distraction is identified as one of the main causes of accidents. These facts have motivated the research community to study cues that can signal driver distractions. Most of the relevant studies have focused on finding differences between normal and distracted driving behaviors [2]–[6]. In same cases, the drivers are considered distracted when they are performing secondary tasks [3], [7], [8]. However, secondary tasks can induce different types of distractions, including cognitive, visual, auditory and manual distractions [9]. Each of them affects the driver's behaviors in particular ways. Furthermore, the driver can become distracted even when they are not engaged in secondary tasks (e.g., thinking and day-dreaming). Therefore, it is essential to study driver distractions by modeling the explicit relationship between driver behaviors and particular distraction types.

This paper studies two of the most common types of driver distraction: cognitive and visual distractions [3]. The study relies on real-world driving data using the UTDrive Platform [10], which is a vehicle equipped with multiple sensors including a microphone array, a frontal camera and a road camera. The car also records various CAN-Bus signals that describe the vehicle activity. Twenty subjects were asked to drive a predefined route performing common secondary tasks that were chosen to induce different level of cognitive and visual distractions. We used subjective evaluations from external observers to assess the perceived cognitive and visual distraction levels of the drivers. The evaluation includes 480 randomly chosen 10 secs videos from our multimodal corpus containing a balanced number of secondary tasks. While perceived distraction scores may differ from the actual distraction levels felt by the drivers, the strong consistency among evaluators in both types of distractions suggests that these labels can be used to train machine learning algorithms.

The first part of the analysis separately considers cognitive and visual distractions. Using the scores from the perceptual evaluations as dependent variables and multimodal features extracted from the segments as independent variables, we propose regression models with elastic net regularization to predict the visual and cognitive distraction levels. We also evaluate binary classifiers where the data is partitioned into high and low distraction classes for both visual and cognitive distractions based on the subjective evaluations. The results reveal high accuracy for both regression and binary classification problems. The study also identifies relevant features characteristic of these two distraction types.

The second part of the analysis proposes a joint cognitive-visual space to characterize driver behaviors. This space reveals four distinctive distraction modes defined by a data driven clustering approach. The analysis of the distribution of the secondary tasks per distraction mode provides insights about the deviation in driving behaviors induced by activities not related to the primary driving task. Using the proposed four distraction modes, drivers' distraction can be characterized in a more general and representative way. We implement machine learning problems that demonstrate the effectiveness of multimodal features in discriminating drivers behaviors associated with each distraction mode. The results indicate that the joint cognitive-visual space provides a general representation to

N. Li and C. Busso are with the Department of Electrical Engineering, University of Texas at Dallas, Dallas, TX, 75080 e-mail: nxl056000@utdallas.edu, Busso@utdallas.edu.

Manuscript received October 16, 2013

identify and categorize the effects induced by secondary tasks. This framework is particularly useful in the assessment of distractions induced by new in-vehicle technology.

The paper is organized as follows. Section II reviews previous studies on driver distraction. Section III describes the multimodal corpus recorded for this study with real driving conditions. Section IV presents the subjective evaluations from external observers to infer the perceived cognitive and visual distractions of the drivers. The section also analyzes the consistency of the evaluations. Section V presents regularized regression models and binary classification evaluations to separately identify cognitive and visual distractions. Section VI proposes a joint cognitive-visual space to represent distraction modes. It also presents classification results to identify the distraction mode of the driving recordings. Section VII concludes the paper with discussion, final remarks and our future research directions.

II. Related Work

One challenge in the study of in-vehicle active safety systems is the wide range of potential distractions that drivers are exposed to. These distractions can be induced by secondary tasks not related to the driving task (i.e., listening to radio, using cellphone, interacting with passengers) or by external events/distractions (i.e., moving pedestrians, advertising boards, road construction). Depending on the distraction, drivers take cognitive, visual, auditorial, and manual resources from the driving task, affecting their situational awareness. This section reviews different driver distraction models, measurements to evaluate the effect of distractions and the current approaches to infer the distraction level of the drivers. We also describe the contribution of this work in the context of previous studies including our own previous work.

A. Multidimensional Distraction Representation

The ideal driver distraction model should consider various factors including drivers differences, traffic and weather conditions, drivers' engagement in secondary tasks, and types of secondary tasks. Given the limitations in acquiring and quantifying these factors, most driver distraction models mainly focus on driver behaviors induced by a limited number of secondary tasks.

Peng et al. [11] categorized driver distractions into visual and non-visual classes to study the drivers' lane keeping ability. Visual distractions were defined as distractions with eye-off-the-road, and non-visual distractions were defined as distractions with eye-on-the-road. This taxonomy was appropriate in their study, given the importance of the drivers' gaze in keeping the lane. Strayer et al. [12] proposed three types of driver distractions: visual, cognitive and manual. They stated that these distractions are not mutually exclusive; performing secondary tasks can induce more than one of the three distraction types. Their study also highlighted the importance of considering the duration and frequency of the task. For example, longer glances are more detrimental for the driver than few short glances [13]. Ranney et al. [9] suggested to consider visual, auditory, manual and cognitive distractions. They also highlighted that these four forms of distractions can have individual or joint effects on the drivers' performance.

Despite the difference among these distraction models, both cognitive and visual distractions are recognized as important aspects to characterize driver behaviors. This observation is consistent with the conclusions in Liang et al. [3], where cognitive and visual distractions were identified as the most dominant factors affecting the drivers' attention. Following this direction, the proposed study explores the use of a cognitive-visual space to represent the effect of secondary tasks on drivers' behaviors.

B. Metrics to Describe Distractions

Studies have proposed different metrics to characterize distractions [13]. While visual distractions can be described with gaze-based features, the estimation of cognitive distractions is an open challenge. Cognitive distraction is determined by the drivers' mental workload. Direct measurements to estimate the brain activity can be difficult and noisy, especially in real driving scenarios. Most studies rely on alternative approaches to measure cognitive distractions including driving performance, secondary task performance, eye glance behavior, physiological measures and subjective assessments.

Some measurements capture behavioral changes associated with increases of mental workload. Young et al. [14] listed common measurements used as driving performance, such as speed, longitudinal control metrics (vehicle following distance) and lateral control metric (lane keeping and steering wheel metrics). With higher cognitive workload, these performance metrics deviate from the ones observed during normal driving conditions. Other studies have proposed secondary task performance metrics, including the number of detected events, the number of incorrect responses to specific questions, and reaction times [15], [16].

Physiological metrics have been used to estimate brain activity. Verwey and Veltman [17] compared different measurement to capture cognitive distractions including *interbeat interval* (IBI), *heart rate* (HR) variability, and the interval between *skin conductance responses* (SCRs). The main limitation of these techniques is the need of intrusive sensors to record the signals from the drivers.

An alternative approach is the use of subjective assessments to estimate cognitive distractions. With self reports, the participants complete a questionnaire to quantify their perceived mental workload level immediately after finishing an experiment. Some of these questionnaires include the *NASA task load index* (NASA-TLX), *driving activity load index* (DALI), *subjective workload assessment technique* (SWAT), *modified Cooper Harper* (MCH) scale and *rating scale mental effort* (RSME) [14]. Another interesting approach is the use of perceptual evaluations from external observers – subjects that did not participate as drivers during the recordings. The underlying assumption is that people observing videos can have similar responses as the ones experienced by the drivers during the recordings. Although the perceived distraction level can be different from the true mental workload, we have discussed the advantages of using this method over other metrics to quantify distraction [18], [19]. Piechulla et al. [20] evaluated a situation-aware phone system with different approaches to assess cognitive workload. One of them was perceptual scores from 20 external observers who were asked to assess the workload induced during a phone call. After watching both the traffic and the driver during a phone conversation, the evaluators annotated how disturbing each phone call was for the driver. Similar approaches were implemented to infer the distraction level of the driver in other related studies [21], [22].

This study relies on human evaluations from external observers to measure the perceived cognitive and visual distraction levels. Our assumption is that the external evaluators with driving experience can identify various relevant cues by watching videos showing both the driver behavior and the road (e.g., primary driving task performance, secondary task performance, eye movement, emotion states, lane keeping).

C. Detecting Driver Distractions

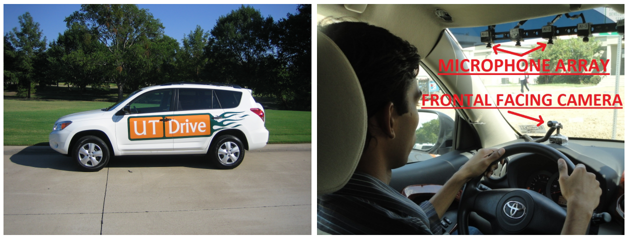
Different studies have focused on modeling the highly non-linear relationship between driver distraction and the aforementioned measurements. Statistic analysis techniques such as ANOVA, correlation analysis and hypothesis test are usually performed to determine whether the measurements are useful for inferring the distraction level of drivers [6], [23]–[26]. Other studies explore machine learning techniques to predict distractive driving behaviors.

Advances in machine learning provide useful tools to capture the highly non-linear dynamics between the drivers and environment. Depending on the measurements under investigation, previous studies have considered various machine learning techniques including *decision tree* [27], *artificial neural network* (ANN) [28], *Adaboost* [29], *support vector machine* (SVM) [2], [29] and *hiddem Markov models* (HMM) [30]. Tseng et al. [27] used decision tree to explore the relationship between driver inattention and accidents. Miyaji et al. [29] compared the performance achieved by SVM and AdaBoost in estimating the drivers' cognitive workload, finding that AdaBoost can achieve higher accuracy. Kutila et al. [2] relied on SVM to detect the driver's visual and cognitive workload. Lee et al. [30] studied driver distraction using 2-state HMMs, where the number of states was determined by a neural network analysis. Tango et al. [28] used an ANN to model driver distraction using drivers' behavioral data collected from simulated recordings.

This study considers six machine learning algorithms for the classification problems: *linear discriminative classifier* (LDC); *k-nearest neighbor classifier* (KNN); *support vector machine* with linear kernel (SVM1); *support vector machine* with quadratic kernel (SVM2); *quadratic discriminative classifier* (QDC); and *Random Under Sampling Boosting* (RUSB). The study aims to find relevant multimodal features signaling cognitive and visual distractions. The framework, features and proposed representation can play a crucial role in designing effective active safety systems.

D. Relation to our Prior Work

We have explored the differences in driving behaviors between normal and secondary task conditions. Li et al. [31] used multimodal features to train binary and multi-class classification to identify drivers engaged in secondary tasks (i.e., is the driver using cellphone?). The classifiers predicted with high accuracy the seven secondary tasks listed in Table I. Since drivers perform a variety of tasks while driving, beyond the ones that we considered, we extended the approach to train regression models to estimate a continuous metric describing the distraction level of the drivers (i.e., how distracted is the driver?) [21]. These regression models can quantify the distraction level induced by secondary tasks over localized segments regardless of the secondary tasks.



(a) UTDrive (b) Sensors

Fig. 1. UTDrive car and sensors placement.

In our previous work, we evaluated the distraction level of drivers using perceptual evaluation without making any distinction between the types of distraction [18], [31]. We realized that when the evaluators were asked to assess just "distraction" they mainly focused on visual distractions. This observation motived us to conduct the perceptual evaluation again, asking our subjects to directly annotate perceived visual and cognitive distractions [19], [32]. Encouraged by the promising results, this study build upon our previous work by proposing a visual-cognitive space to evaluate and estimate driver distractions. By detecting general mode of distractions, we can evaluate the detrimental effect of any secondary task on driving behaviors. The contributions of this study are:

- We explicitly study driver behaviors in terms of both perceived visual and cognitive distractions, which are estimated with perceptual evaluations.
- We propose regression models and binary classifiers to predict perceived cognitive and visual distractions using multimodal features extracted from noninvasive sensors.
- We propose a joint visual-cognitive space to better represent distractions introduced by secondary tasks.

III. Data Collection

This study relies on a multimodal corpus collected using the UTDrive platform – a 2006 Toyota RAV4 equipped with various sensors (Fig. 1 (a)). The UTDrive platform has a camera on the dashboard facing the driver's face (640 × 480, 30 fps) and a camera facing the road (320 × 240, 15 fps). It has a microphone array and a pressure sensor on the gas pedal. The *controller area network-bus* (CAN-Bus) data is extracted and recorded simultaneously with the video and audio signals in a Dewetron computer placed behind the driver. Details about the car and its unique features are described in Angkititrakul et al. [33].

TABLE I
SECONDARY TASKS IN THE CORPUS. THE TASKS PHONE AND GPS WERE SPLIT, SINCE WE OBSERVED DIFFERENT DRIVER BEHAVIORS WHILE OPERATING AND USING THE DEVICES.

Task Name	**Description**	**Duration [s]**		**Mean Speed [km/hr]**	
		Lap 1	**Lap 2**	**Lap 1**	**Lap 2**
Radio (red-route)	Driver tunes the radio to predetermined stations	1748	1635	50.56	54.04
GPS-Operating (green-route)	Driver inputs predetermined address into GPS	1113	975	48.30	55.46
GPS-Following (green-route)	Driver follows the GPS instruction to the destination	3286	3050	35.13	37.09
Phone-Operating (blue-route)	Driver dials the airline automatic flight information system using a cellphone	476	478	51.76	53.46
Phone-Talking (blue-route)	Driver interacts with the flight information system to retrieve flight information	2403	2546	40.76	37.81
Picture (orange-route)	Driver describes the A4 size pictures shown by the passenger to reflect the distraction caused by road signs and billboard	1564	1573	51.79	52.34
Conversation (black-route)	Driver discusses the driving experience with the passengers and answers general questions	1648	1618	43.75	45.37

Twenty subjects (10 male and 10 female) with valid driver license participated in the data collection. The average and standard deviation of the participants' age are 25 and 7, respectively. A predetermined 5.6-miles city route was selected for the recording, as shown in Fig. 2. The average speed limit of the selected route is approximately 37 miles/hr (53.5 km/hr). The participants were asked to drive along the same route twice. During the first lap, the drivers were asked to perform the secondary tasks described in Table I in sequential order: operating a radio (*Radio*), operating and following a *global positioning system* (GPS) (*GPS - Operating* and *GPS - Following*), operating and talking on a cellphone (*Phone - Operating* and *Phone - Talking*), describing pictures (*Pictures*) and conversation with a fellow passenger (*Conversation*). These activities are secondary tasks that are commonly conducted by drivers in real driving scenarios. They cover various dimensions of distraction including visual (e.g., *Radio*, *GPS - Operating*, *Phone - Operating*, *Pictures*), manual (e.g., *Radio*, *GPS - Operating*, *Phone - Operating*), cognitive (e.g., *Phone - Talking*, *Conversation*), and auditory (e.g., *Conversation*, *Radio*) distractions. Table I gives the details about the secondary tasks, including the duration and average speed for each task.

In the second lap, the participants drive along the same route without performing any secondary task. By fixing the order of the tasks over predefined route segments, we can collect a reliable recording baseline for normal driving behavior, in which most of the other variables are kept fixed (e.g., traffic signal, route curves, speed limit). With this controlled recording, we can study the differences in driving behaviors during tasks and normal conditions. Therefore, the observed differences can be mainly associated with the behaviors induced by secondary tasks. A detailed description of the protocol and recording settings can be found in Li et al. [31].

IV. DRIVER DISTRACTION ASSESSMENT USING PERCEPTUAL EVALUATIONS

This study relies on perceptual evaluations from external evaluators to assess visual and cognitive distractions induced by different secondary tasks (Table I). The external observers assessed the distraction level of the drivers after watching short videos showing the driver and the road (Sec. IV-A). To unify their understanding of cognitive and visual distractions, we carefully instructed the evaluators with their definitions. We follow the description given by Ranney et al. [9]. Visual distraction is defined as "eye-off-the-road" – drivers looking away from the roadway. The evaluators were asked to rate the visual distraction level based on the glance behavior of the drivers. The road camera was included to help the evaluators to assess whether the observed head motion or eye glancing behaviors were related to the primary driving task. Cognitive distraction is defined as "mind-off-the-road" – drivers being lost/busy in thought. For cognitive distraction, the evaluators were asked to rate the videos based on their own judgment. However, we highlighted that facial expressions (stress level, eye pupil size, eye movements), secondary task performance (talking speed, phone dialing speed) and driving performance (vehicle in-lane position, driving speed, distance to front vehicle) could be used to assess the cognitive distraction level. The idea behind this methodology is that subjects with driving experience are able to extract information perceived from multiple modalities (e.g., frontal camera, road camera and audio), process the stimuli, infer the perceived visual and cognitive distractions, and provide consistent assessments that reflect the relative distraction levels experienced by the drivers

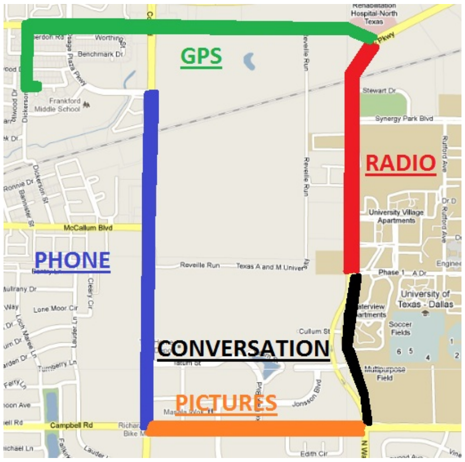


Fig. 2. Route used for the study. The subjects drove this route twice (5.6 miles): first, performing a series of secondary tasks starting with *Radio* and ending with *Conversation* (Table I); and then, driving normally without getting involved in any secondary task.

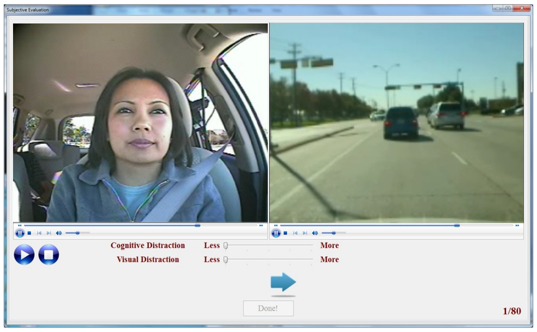


Fig. 3. GUI used to separately evaluate the perceived visual and cognitive distractions.

during the videos.

Notice that in our previous study, we compared perceptual evaluations from external observers with two commonly used non-intrusive approaches for labeling driver's distraction: self-evaluation and eye glance metrics [18], [19]. The analysis suggests important advantages of using human evaluation from external observers over other alternative approaches. When comparing to self-evaluation, multiple external evaluators can provide reliable scores for short videos that are closely related to the distraction level in specific scenarios. When comparing to eye glance metrics, external observers can take advantage of the driving dynamics perceived from multiple modalities (e.g., facial expression, road condition) and provide more reliable assessments than objective metrics that consider only eye glance behaviors. The study also showed that the perceived distraction scores were reliable and consistent across evaluators.

A. Subjective Evaluation

The database consists of over 12 hours of real driving recordings. To limit the number of perceptual evaluations, only a subset of the corpus is used for this study. The videos consist of 10-second non-overlapped recordings with synchronized information from different modalities, including frontal camera, road camera and microphone (See Fig. 3). For every driver, three videos are extracted from each of the 8 driving conditions (seven secondary tasks and normal conditions). Altogether, 480 videos are subjectively evaluated (20 drivers × 8 driving conditions × 3 segments). Notice that the selected videos include a balanced set of driving conditions (60 videos for each driving condition).

Thirty college students from different disciplines were asked to evaluate both the perceived visual and cognitive distractions after watching the audiovisual stimuli. All the subjects had driving experience at the time of the evaluation. To ensure consistency in the evaluations and to reduce the learning curve effect, the visual and cognitive distraction evaluations are separately conducted. Each external observer first evaluated 80 videos (10 drivers × 8 driving conditions) for visual distraction (around 18 minutes), and then evaluated 80 different videos (different 10 drivers × 8 driving conditions) for cognitive distraction (around 25 minutes, since it takes longer for evaluators to judge the cognitive distraction level). We suggested a 30 minutes break between the two evaluations.

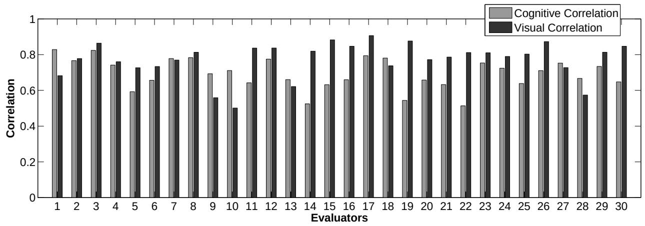


Fig. 4. Correlation for cognitive and visual distraction scores for each external evaluator.

Overall, the 30 external observers generated 4800 evaluations: 2400 for visual (30 × 80) and 2400 for cognitive (30 × 80) distractions. As a result, each of the 480 video segments is evaluated by ten different external observers, five for perceived visual distraction and five for perceived cognitive distraction. Figure 3 shows the *graphical user interface* (GUI) used for the subjective evaluation. The evaluator gives a continuous value on a scale from 0 (least distracted) to 1 (most distracted) by adjusting the sliding bar.

B. Inter Evaluator Agreement

The inter evaluator agreement analysis is conducted to measure the degree of consistency among different evaluators. High consistency among different evaluators indicates strong agreement, suggesting reliable perceived cognitive and visual scores. We used leave-one-out correlation to measure the similarity between the scores across evaluators. In this approach, the correlation is separately calculated for the perceived visual and cognitive scores. The evaluation from one subject is compared with the average evaluation from the other four evaluators. The average correlations for the perceived visual and cognitive distractions are $\hat{\rho}^{vis} = 0.77$ and $\hat{\rho}^{cog} = 0.69$, respectively. This result suggests strong correlation in both visual and cognitive distraction scores. The lower correlation for the perceived cognitive distraction scores highlights the challenging task of predicting the drivers' cognitive workload. However, the correlation still represents a strong positive relationship between the scores provided by the external observers. Figure 4 shows the correlation calculated for each evaluator. The correlation for both cognitive and visual distractions is always higher than $\rho = 0.5$ for each individual. Notice that inter-evaluator agreement can be considered as an upper bound for the regression models presented in Section V.

C. Perceptual Evaluation Analysis

Fig. 5 provides the average and standard deviation of the subjective evaluation for the seven secondary tasks (Table I) and normal condition. The results suggest that the evaluators perceived *Radio*, *GPS - Operating*, *Phone - Operating* and *Pictures* as the most distracting tasks both visually and cognitively. Although global statistics for some of the tasks are similar for cognitive and visual distractions, a deeper analysis illustrates the benefits of using a two-dimensional space to characterize driver distractions. Fig. 6 provides a scattering plot of the perceived distraction scores. The figure shows samples distributed over most of the 2D space. The only

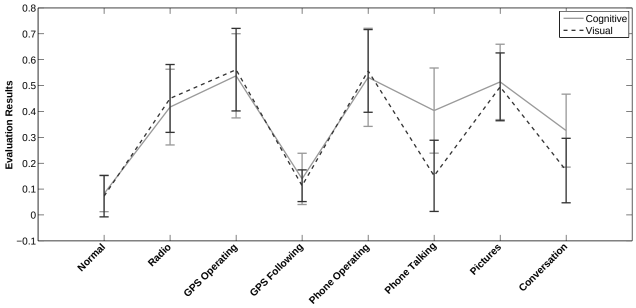


Fig. 5. Mean and standard deviation of the subjective evaluation scores. *Phone - Talking* and *Conversation* present the highest differences between cognitive and visual evaluations.

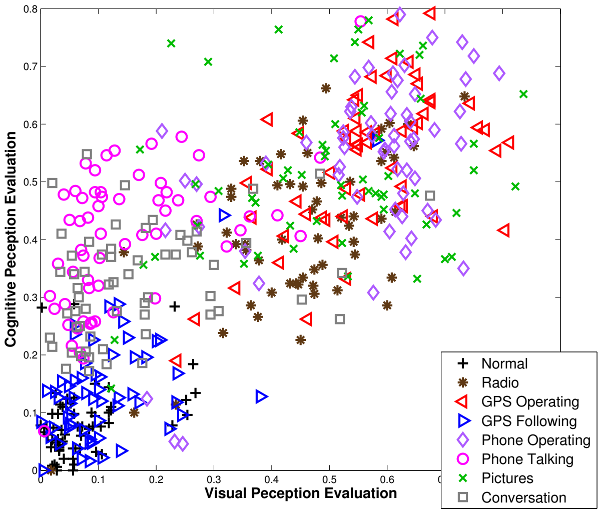


Fig. 6. Scattering plot of the subjective evaluation across secondary tasks in the visual-cognitive space.

exception is the area corresponding to low cognitive distraction and high visual distraction (bottom-right quadrant). This result is expected, since some of the visual secondary tasks also induce cognitive workload (e.g., tuning radio, describing pictures, operating a phone). However, the cognitive tasks such as *Phone - Talking* and *Conversation* do not necessary increase visual demand (these tasks present the greatest differences between visual and cognitive distraction scores – Fig. 5). This finding agrees with the *100-Car Naturalistic Driving Study*, which indicated that the secondary tasks *Interacting with Passenger* and *Talking/Listening on Phones* are almost exclusively cognitive in nature [34]. Distraction defined in the visual-cognitive space is further discussed in Sec. VI.

V. Predicting Visual and Cognitive Distractions

This section explores machine learning algorithms to predict the visual and cognitive distractions of the drivers. The systems are trained using the perceptual evaluations as ground truth. After describing the features (Sec. V-A), this section discusses regression (Sec. V-B) and binary classification (Sec. V-C) analyses.

TABLE II

Multimodal Features. LLFs are time series signals over which we estimate statistics. HLFs are single values derived from LLFs (*CAN* = CAN-Bus signal; *MI* = microphone; *RC* = road camera; *DC* = driver camera)

	Low Level Features (LLFs)	
CAN	Vehicle Speed (Speed)	Brake Pressure (Brake)
	Steering Wheel Angle (Steering)	Acceleration (Acceleration)
	Steering Wheel Jitter (Jitter)	Gas Pedal Pressure
	Brake Pedal Pressure	
MI	Energy	
	10 Gammatone Filter Responses	(Audio GT 1-10)
RC	Road Optical Flow	
	Road Intensity	
DC	Head Yaw Angle (Yaw)	Chin Raiser (AU17)
	Head Pitch Angle (Pitch)	Lip Stretcher (AU20)
	Head Roll Angle (Roll)	Cheek Raiser (AU6)
	Inner Brow Raiser (AU1)	Lip Tightener (AU7)
	Outer Brow Raise (AU2)	Lip Puckerer (AU18)
	Brow Lowerer (AU4)	Lip Tightener (AU23)
	Upper Lid Raiser (AU5)	Lip Pressor (AU24)
	Nose Wrinkler (AU9)	Lips Part (AU25)
	Upper Lip Raiser (AU10)	Jaw Drop (AU26)
	Lip Corner Puller (AU12)	Lip Suck (AU28)
	Dimpler (AU14)	Eye Openness (AU45)
	Lip Corner Depressor (AU15)	
	High Level Gaze Features	
	Eyes-Off-the-Road Duration (EOR Dur.)	
	Eyes-Off-the-Road Frequency (EOR Freq.)	
	Longest Eyes-Off-the-Road Duration (LEOR Dur.)	
	Eye Blink Frequency (Blink Freq.)	

A. Multimodal Features

In our previous studies, we have shown that valuable information signaling drivers' distraction can be extracted from nonintrusive sensors [6], [31], [32]. In particular, we considered features extracted from the *controller area network* (CAN)-Bus, a camera facing the drivers and a microphone array. Here, we included the camera facing the road and extended the feature set derived from these modalities (Table II). Our approach consists in estimating frame-by-frame signals from the 10-sec videos, referred to as *low level features* (LLFs). Then, eight statistics or *high level features* (HLFs) are estimated over the segments for each LLF. The statistics are the average (Mean), standard deviation (STD), maximum (Max), minimum (Min), range (Ran), interquartile range (IQR), skewness (Ske) and kurtosis (Kur).

CAN-Bus Signal (CAN): CAN-Bus provides useful information about the vehicle's activity including vehicle speed, steering wheel angle, brake value, and RPM acceleration [10], [33], [35]. This study uses the steering wheel angle, vehicle speed, brake value and RPM acceleration extracted from the CAN-Bus signal as LLFs. It also uses data from pressure sensors on the gas and brake pedals, which provides the pedal pressure. We also estimated the steering wheel variance over 5 sec windows which is referred to as steering wheel jitter. This feature describes small corrections in the steering wheel made by drivers to keep the line. These seven low level CAN-Bus features are used to represent the vehicle activity.

Microphone Array (MI): The microphone provides another modality that is particularly useful to capture driver's activities producing sound (e.g., *Conversation*, *Phone-Operating* and *Radio*). We estimate the RMS energy of the central microphone. In addition, we calculate the gammatone filterbank to extract auditory features at different frequencies, which can better represent low, middle and high frequency sounds produced by the engine, radio, and passengers. Gammatone filters are popular linear filters that approximate the human auditory system. Notice that more filters are assigned to lower frequencies, as the bandwidth of the filters increases for higher frequencies. We use 10 gammatone filters between 100Hz and 25KHz using the implementation defined by Slaney [36].

Camera Facing the Road (RC):

The road camera captures the street condition during the data collection. The required attention level depends on the traffic condition and driving maneuver. Horberry et al. [37] showed that drivers reduced the vehicle speed when facing complex road scenarios. Brookhuis and de Waard [38] found evidences to support this compensatory effect, in which drivers attempt to maintain a reasonably stable level of task difficulty during a journey. They achieve this goal by reducing the frequency of activities such as checking the mirrors. These findings suggest that road information provides useful information to understand the drivers cognitive and visual demand. By adding road features, in addition to other metrics, we expect to exploit these relationships. In this study, we extracted two LLFs: optical flow energy and image intensity. While optical flow emphasizes movement in the road scene, the image intensity captures the global characteristic of the road scene (i.e., intersection, highway, etc).

Camera Facing the Driver (DC): The camera facing the driver is an important sensor for this study. From the video, we automatically extract LLFs describing head orientation, facial expressions and eye-glance behaviors. We rely on the *computer expression recognition toolbox* (CERT) [39]. CERT can estimate these facial features under various illumination conditions, which is important for this study. Notice that the toolkit has been used to detect drowsy driver [40]. A primary advantage of CERT is that the estimation is done frame-by-frame, so errors do not propagate across frames. CERT estimates the head position by first classifying the face pose into discrete head orientations for yaw and pitch movements. Then, the output of the classifiers are used to build a standard linear regression model which gives the precise head orientation (yaw, pitch and roll movements) [41]. In our previous work, features extracted from the drivers' head pose were useful in binary classifiers, aiming to recognize drivers engaged in secondary tasks [6], [31] (i.e., task versus normal conditions). Therefore, we expect that they will be also useful for detecting visual and cognitive distractions.

The facial expressions are estimated in terms of *action units* (AUs). AUs are the building blocks of the *facial action coding system* (FACS) which was developed to describe the face appearance [42]. We use 20 AUs plus 3 head pose angles estimated by CERT (see Table II). Notice that facial expressions captures drivers' behaviors such as smiling, frowning, blinking and speaking, which we hypothesize will be useful to characterize cognitive distractions [32].

Gaze metrics are important features for detecting people's attentions. Therefore, they have been used in studies on driver behaviors [16], [43]. Detecting the actual gaze is extremely difficult given the recording conditions in the car, with adverse illumination. Therefore, eye-glance behaviors are estimated from the drivers' head position (we acknowledge that this is an approximation). For each driver, we estimate a rectangular reference field centered at the road using his/her normal driving behavior. The width and height of the rectangle is defined by two standard deviations of the head yaw and pitch angles, respectively. We consider that the driver has his/her eye-off-the-road for a given frame when the estimated angles lie outside this reference field. Using this information, we estimate the following statistics for each of the 10-sec videos: the eye-off-the-road duration (EOR Dur.) (i.e., number of eye-off-the-road frames), eye-off-the-road frequency (EOR Freq.)(i.e., number of transition from eye-on-the-road frame to eye-off-the-road frames), and the longest eye-off-the-road duration (LEOR Dur.) (i.e., the duration of the longest glance). These features are closely related to driver distraction [13], [19]. We use the estimation of eye closeness (AU45) to calculate the drivers' eye blink actions (high values indicate eyes that are closing). The mean and standard deviation of AU45 is calculated for each driver using the observation from the normal driving conditions. On average, people blink 10 times per minute with 300ms blink duration when the eyes are focus on objects (5% of the time) [44]. Therefore, we set the threshold for blink detection as the mean plus two standard deviations to match this statistic. Finally, we estimate for each of the 10-sec video the blink frequency (i.e., number of transition from open-eyes frames to close-eyes frames). Unlike LLFs, these four parameters are high level gaze features estimated over the entire video, so no statistics are estimated over them (last four rows of Table II).

We estimate a 348D feature vector (43 LLFs × 8 HLFs + 4 gaze features). In frames with adverse illumination conditions or occlusion by the driver's hands, CERT fails to capture the driver's face producing missing values. If the number of frames with missing values is more than one third of the total number of frames in the video, we discarded the videos (85 out of 480 videos). Otherwise, the missing values are interpolated with the adjacent values. This interpolation introduces a delay that can affect real-time implementation of the proposed methods.

B. *Regularized Regression*

The first part of the analysis consists in building linear regression models to predict the perceived visual and cognitive distraction levels. The multimodal features are the independent variables and the visual and cognitive distraction scores are the dependent variables. We reduce the high dimensional feature vector using elastic net regularization. Elastic net is a regularization technique that combines LASSO (L^1) and ridge (L^2) optimization methods. By considering a penalty term, it reduces the number of coefficients, and, therefore, the selected features. Given N observations $\{(x_1, y_1)\dots(x_N, y_N)\}$, elastic net solves the following equation for the intercept and coefficient vector:

$$min_{\beta_0,\overline{\beta}}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\overline{\beta}\right)+\lambda P_\alpha(\overline{\beta})\right\} \quad (1)$$

TABLE III
REGRESSION RESULTS

	Visual Distraction Regression		Cognitive Distraction Regression	
	Training	**Testing**	**Training**	**Testing**
Correlation	0.753	0.704	0.679	0.645
# of Feat Mean	19.3		21.2	
# of Feat Std	2.41		1.94	

where

$$P_\alpha(\overline{\beta}) = \sum_{j=1}^{n}(\frac{(1-\alpha)}{2}\beta_j^2 + \alpha\,|\beta_j|). \qquad (2)$$

The parameter α is strictly between 0 and 1, and λ is a nonnegative regularization parameter. Both parameters determine the number of features in the regression model. When α approaches 0, the model becomes a ridge regression model which increases the number of features. When α approaches 1, the model becomes a LASSO regression model where few coefficients are set to nonzero values, especially when the independent variables are highly correlated. For a fixed value of α, increasing the value of λ decreases the number of nonzero components. The maximum number of independent variables in the regression model is limited to 30, given that we only consider 395 videos. This constrain limits the complexity of the regression model while satisfying the suggested ratio between number of independent variables and sample size [45]. After training the model, the independent variables with non-zero coefficients correspond to features highly related to dependent variables – the perceived visual and cognitive scores.

In addition to the training and testing sets, we define a development set to estimate the parameters λ. To maximize the usage of the database, we create these partitions using two-layer driver independent cross-validation approach (i.e., all the videos from one driver are exclusively in one of the partitions). First, we use a 20 fold cross-validation approach, in which the videos from one driver are used as testing set. Then, with the videos of the remaining drivers, we define the training and development set with a second 19-fold cross-validation approach (20 × 19 = 380 folds). We use the videos from one driver as development set, and the videos from the remaining 18 drivers as training set. In each fold, we fix α at 0.5 to balance the dependency on LASSO and ridge optimization methods, and we increase the value for λ such that the number of independent variables changes from 5 to 30. For a given driver, we identify the best performance observed over the development set across the corresponding 19 folds. We set the value of λ such that the model has approximately the optimum number of features, as observed in the development set. Finally, we merge the training and development sets and build the regression model. We evaluate the performance of the system with the testing set, which is neither used for building the models nor setting the value of λ.

We evaluate the performance of the regression models using the correlation between the predicted distraction level and the perceived distraction annotations. Table III reports the average correlation across 20 folds. The results are estimated over the testing partitions and the training+development sets. The regression models provide high correlation scores (ρ^{vis}=0.704 and ρ^{cog}=0.645). The close performance observed when the models are evaluated in the training and testing sets suggests that the regression models generalize to the behaviors displayed by drivers whose videos were not included to train the models. The table also provides the mean and standard deviation of the number of features included in the model. The low standard deviations suggest that the proposed training approach produces stable models.

TABLE IV
SELECTED FEATURES IN THE REGRESSION MODELS.

Visual Regression	Cognitive Regression
Speed IQR	Inner Brow Raiser (AU1) Max
Brake Press Sensor STD	Outer Brow Raiser (AU2) Max*
Yaw Mean*	Outer Brow Raiser (AU2) Range*
Outer Brow Raiser (AU2) Max*	Lip Corner Puller (AU12) Range*
Outer Brow Raiser (AU2) Range*	Chin Raiser (AU17) STD
Lid Tightener (AU7) IQR*	Lid Tightener (AU7) IQR*
Lip Tightener (AU23) IQR*	Lip Tightener (AU23) IQR*
Jaw Drop (AU26) Max*	Jaw Drop (AU26) Max*
Blink Freq.*	Blink Freq.*
Brake IQR	EOR Dur.
Audio GT6 Mean	Steering IQR*
Steering IQR*	Yaw Mean*
Lip Pressor (AU24) Range*	Lip Pressor (AU24) Range*
Lip Corner Puller (AU12) Range*	Speed Min
Lip Puckerer (AU18) Max	Roll Mean
	Lip Tightener (AU23) Max
	Roll Kurtosis
	Lip Suck (AU28) Max

High level features selected in both distraction models are marked with (*).

The regression models provide a systematic way to identify features related to visual and cognitive distractions. Notice that the independent variables considered in the models vary across folds. For consistency, we identify features that are selected in at least 75% of the 20 folds (first layer of cross-validation). Table IV lists the 15 features selected for visual distraction, and 18 features selected for cognitive distractions. There are certain features that are included in both groups – highlighted with (*). There are also unique variables describing visual and cognitive distractions.

The independent variables for visual distraction models includes features extracted from: CAN-Bus signal (Speed IQR, Brake IQR, Brake Press Sensor STD), microphone array (Audio GT6 Mean) and camera facing the driver (Lip Puckerer (AU18) Max). The result that Speed IQR, Brake IQR, Brake Press Sensor STD are selected for visual distraction is consistent with previous findngs that visual distraction has a greater effect on lateral control measures [14].

The features for cognitive distraction models come from facial features (Inner Brow Raiser (AU1) Max, Chin Raiser (AU17) STD, EOR Dur., Roll Mean, Lip Tightener (AU23) Max, Roll Kurtosis, Lip Suck (AU28) Max) and CAN-Bus signal (Speed Min). We found in our previous work that head roll movements was useful to recognize drivers using cellphones [31]. Given the detrimental effect of phone talking on cognitive distractions, it is interesting to observe that the features related to head roll movement are included (Roll Mean). The feature Eye-Off-the-Road Duration, which was the most frequently selected feature for binary classification

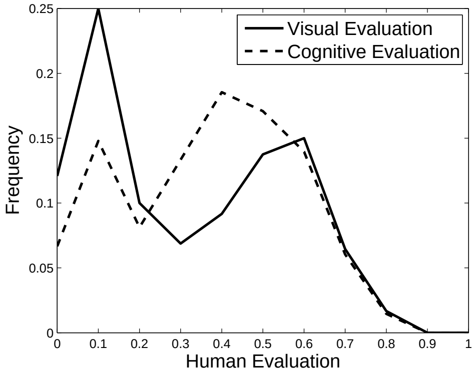


Fig. 7. Distribution of perceived visual and cognitive perceptual scores. The bimodal distributions define the classes with low and high distraction levels.

between normal and task driving conditions [31], is only selected for the cognitive regression models. Notice that certain eye-off-the-road behaviors such as mirror checking are characteristic primary driving tasks [46]. While these actions induce higher cognitive load, they might not be perceived as visual distractions by the evaluators.

C. Binary Classification

An alternative approach to identify driving behaviors is to classify recordings with low or high level of distractions. This section addresses this problem with two separate binary classifiers that recognize low versus high level of visual or cognitive distractions. Figure 7 shows the distributions of the scores from the perceptual evaluation across secondary tasks. For both cases, we observe a bimodal distribution representing low and high level of distractions. This pattern may be the result of considering the selected secondary tasks. The figure clearly suggests that the two modes are separated at $t_{cog} = 0.2$ for cognitive distractions, and $t_{vis} = 0.3$ for visual distraction. These thresholds are selected to create the two binary classification problems. For cognitive distractions, we have 105 videos with low level of distraction, and 290 videos with high level of distraction. For visual distractions, the two classes have 214 (low) and 181 (high) videos, respectively.

We use *forward feature selection* (FFS) based on Mahalanobis distance to reduce the feature set. Given a desired number of features n, FFS selects the feature set that maximizes the sum of the Mahalanobis distances between classes. We determine the value of n using the two-layer cross-validation approach discussed in Sec. V-B. The videos are split in driver independent partitions for the training (18 drivers), development (1 driver) and testing (1 driver) sets. For a given driver whose videos are in the testing set, we use a 19-fold cross-validation to define the training and development sets using the videos from the remaining drivers. We build the classifiers using the training set by changing the number of features from 1 to 30. Then, we set n equal to the average number of features that maximizes the performance on the development set across the 19 folds. After we define the number of features for a given driver in the testing set, we build the classifiers using both the training and development videos. Finally, we report the results on the testing set over the 20 folds (Table V).

TABLE V

BINARY CLASSIFICATION OF LOW AND HIGH DISTRACTION LEVELS FOR VISUAL AND COGNITIVE DISTRACTIONS. THE TABLE REPORTS THE AVERAGE PRECISION (**P**), AVERAGE RECALL (**R**) AND F-SCORE (**F**).

	Visual Distraction			**Cognitive Distraction**		
	P [%]	**R** [%]	**F**	**P** [%]	**R** [%]	**F**
LDC	77.5	77.0	0.772	79.2	79.5	0.794
KNN	73.3	71.3	0.723	69.8	66.5	0.681
SVM1	77.5	77.0	0.772	79.4	78.6	0.790
SVM2	76.9	76.6	0.767	69.4	64.8	0.670
QDC	76.3	76.4	0.764	73.2	76.2	0.747
RUSB	73.4	72.9	0.731	77.5	80.9	0.791

The selection process does not maximize the performance of a particular classifier, so we use the feature set to compare different machine learning algorithms. We consider six classifiers: *Linear Bayes Normal Classifier* (LDC); *K-Nearest Neighbor* (KNN) Classifier; *Support Vector Classifier* (SVM1) with Linear Kernel; *Support Vector Classifier* (SVM2) with Quadratic Kernel; *Quadratic Bayes Normal Classifier* (QDC); and *Random Under Sampling Boosting* (RUSB). Since the data is not balanced, we include RUSB that is designed for unbalanced classification problems [47]. While all other classifiers use the selected features, RUSB uses boosting over the entire feature set.

We use the average *precision* (P), average *recall* (R) and *F-score* (F) to evaluate the classifiers. For each distraction type, we estimate the precision rate for low and high classes – fraction of retrieved samples for one class that are relevant. Then, we estimate and report the average precision for low and high classes. Likewise, we estimate the recall rate for each class – fraction of relevant samples that are correctly classified. We report the average recall for low and high classes. With these values, we calculate the F-score using (3), which is used as a single measurement to evaluate the performance of the classifiers. This metric is not affected by unbalanced sets.

$$F = \frac{2PR}{P + R} \tag{3}$$

Table V gives the performance of the binary classifiers. LDC provides the best F-score on visual (0.772) and cognitive (0.794) classification problems. A similar performance is achieved by SVM1. For cognitive distractions where the data is more unbalanced (1:2.8), we observe that RUSB achieves the highest recall rate (80.9%). The selected features in this experiment can be used to discriminate between low and high distraction levels. They vary across folds due to changes in the training/development data. For visual distraction, the mean and standard deviation of the number of features used for LDC are μ_{LDC}^{Vis}=9.07 and σ_{LDC}^{Vis}=1.88, respectively. For cognitive distraction, these two statistics are μ_{LDC}^{Cog}=11.06 and σ_{LDC}^{Cog}=1.58.

Table VI lists in order the ten most frequently selected features for visual and cognitive classification problems. Many of the selected feature are consistent with the features selected in the corresponding regression model (marked with $\star$). The

TABLE VI
THE TEN MOST FREQUENTLY SELECTED FEATURES FOR VISUAL AND COGNITIVE CLASSIFICATION PROBLEMS.

Visual Classification	Cognitive Classification
Lip Tightener (AU23) IQR⋆	Audio GT8 Skewness
Jaw Drop (AU26) Max⋆	Road Intensity STD
Yaw Mean⋆	Lip Corner Depressor (AU15) STD
Brake IQR⋆	EOR Dur.⋆
Road Optical Flow Mean	Audio GT8 Kurtosis
Outer Brow Raiser (AU2) Max⋆	Outer Brow Raiser (AU2) Max⋆
Audio GT6 Range	Pitch Kurtosis
Pitch Kurtosis	Lip Tightener (AU23) IQR⋆
Blink Freq.⋆	Speed IQR
Brake Press Sensor Min	Lip Corner Depressor (AU15) Kurtosis

Features also selected in the regression models (Table IV) are marked with ⋆.

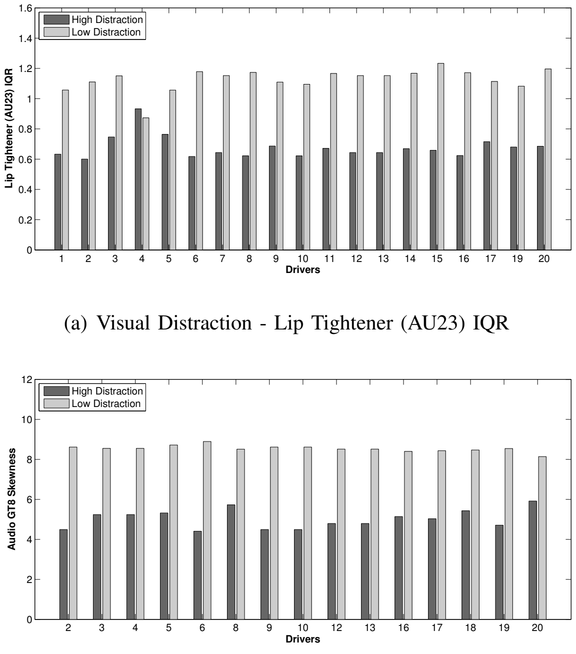
table includes features from all the modalities for both visual and cognitive binary classifications (road camera, face camera, microphone, and CAN-Bus). Unlike Table IV, the overlap between the most selected features for visual and cognitive classification tasks has only three common features: Pitch Kurtosis, Lip Tightener (AU23) IQR and Outer Brow Raiser (AU2) Max.

There are clear differences between visual and cognitive distractions. For visual distraction, the selected feature set supports the strong relationship between glance related features and visual distractions (Yaw Mean, Road Optical Flow Mean, Outer Brow Raiser (AU2) Max, Pitch Kurtosis and Blink Freq.). It also highlights the discriminative power of drivers speed control features (Brake IQR, Brake Press Sensor Min). This result agrees with Young et al. [14], which indicates that visual distraction affects lateral control measurements. For cognitive distractions, three AU features are selected for the binary classification: AU5, AU15, and AU18 (see Table II). These AUs describe the movement of brows and lips. These results further indicate the benefits of using facial features to detect perceived cognitive distractions.

We compare the mean values per driver of the most frequently selected feature for the binary classification problems. As listed in Table VI, these features are Lip Tightener (AU23) IQR for visual distraction, and Audio GT8 Skewness for cognitive distraction. For some of the drivers, we have unbalanced number of videos between the two classes. To estimate reliable mean values, we only consider drivers when at least six of their videos were assigned to each of the two classes. This threshold discards one driver for visual distraction, and five for cognitive distractions. For visual distraction, Figure 8(a) shows that the patterns are very consistent across the drivers (except for driver four). Lower values of interquartile range (IQR) of AU23 are observed for visual distractions. For cognitive distraction, Figure 8(b) shows that the feature has consistently lower values for the high cognitive class.

VI. CLUSTERING ANALYSIS

The previous section considers the detection of visual and cognitive distractions as two independent problems. However, addressing each distraction dimension separately provides limited insights to evaluate the overall drivers' distraction. As mentioned in Sec. II-A, the driver's behaviors can reflect the result of the joint interaction of multiple distraction sources. This section considers the detection of driver's distraction using a joint visual-cognitive space.



(a) Visual Distraction - Lip Tightener (AU23) IQR

(b) Cognitive Distraction - Audio GT8 Skewness

Fig. 8. Average values per driver for the most selected features by binary classifiers for visual and cognitive distractions (low versus high).

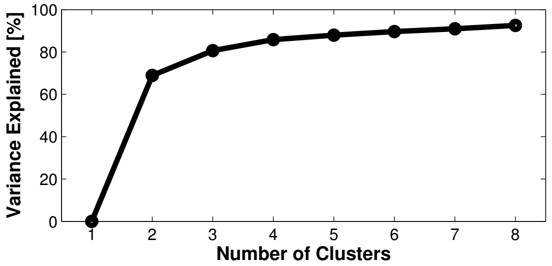


Fig. 9. *Elbow* method used to define the number of clusters for the visual-cognitive space. After four clusters, the percentage of variance explained does not significantly increase.

A. K-means Clustering

Figure 6 reveals that the perceived scores are distributed across the entire visual-cognitive space (except the lower right corner of the plot – see discussion in Sec. IV-C). From this plot, we aim to automatically define modes of distraction that will be useful to characterize driver behaviors. We implemented the unsupervised k-means cluster algorithm to identify the distraction modes. We use the *elbow* method to select an appropriate number of clusters. The elbow method looks at the percentage of variance explained in the data set as the number of cluster increases. The final number of cluster is selected for the case in which adding a new cluster does not significantly affect the percentage of the variance explained by individual clusters. Figure 9 shows the results for the perceived visual and

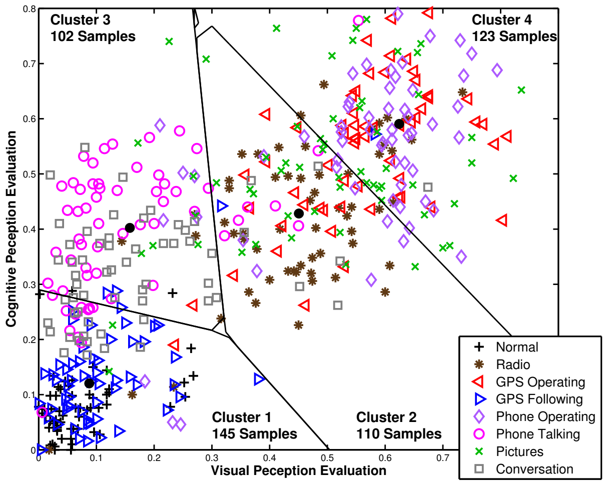
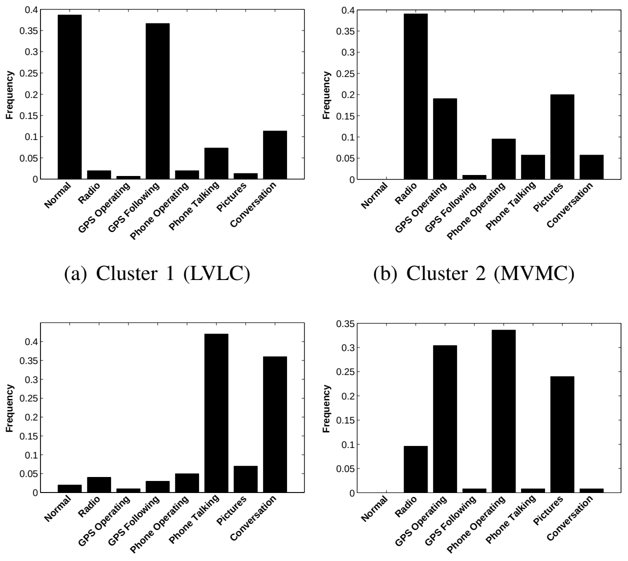


Fig. 10. Four distraction modes defined by the k-means cluster algorithm. The figure gives the number of samples assigned to each cluster.



(a) Cluster 1 (LVLC)

(b) Cluster 2 (MVMC)

(c) Cluster 3 (LVMC)

(d) Cluster 4 (HVHC)

Fig. 11. Distribution of secondary tasks for each distraction mode.

cognitive scores, which indicate that four clusters is sufficient to represent the data.

One drawback of the k-means approach is that the outcome heavily depends on the center of the initial clusters. Instead of using random seed to initialize the clusters, we use the *genetic algorithm* (GA) clustering technique to estimate a reasonable initial cluster [48]. The clusters' centers are used to initialize the k-means algorithm. We evaluate the k-means algorithm multiple times, and select the clusters with minimum total in-class distance. Figure 10 shows the resulting clusters for this case. Using the labels provided in the figure, the clusters represent important distraction modes:

- Cluster 1: Low visual, low cognitive distractions (LVLC)
- Cluster 2: Medium visual, medium cognitive distractions (MVMC)
- Cluster 3: Low visual, medium cognitive distractions (LVMC)
- Cluster 4: High visual, high cognitive distractions (HVHC)

Figure 11 provides the distribution of secondary tasks for each cluster. In addition to *Normal*, Figure 11(a) shows that cluster 1 mainly consists of *GPS-Following*, which is perceived as the less distracting secondary task. Cluster 2 includes samples from most of the secondary tasks (Fig. 11(b)). Cluster 3 includes cases in which cognitive distractions are not necessary associated with visual distractions (i.e., *Phone - Talking* and *Conversation* – Fig. 11(c)). Typical "mind-off-the-road" scenarios are included in this distraction mode where the driver is not visually distracted but his/her cognitive workload is high. Cluster 4 includes visually intensive secondary tasks such as *Phone - Operating*, *GPS - Operating* and *Pictures* (Fig. 11(d)).

B. Classification of Distraction Modes

We expect that secondary tasks inducing distraction levels in clusters 3 and 4 will be more detrimental to the primary driving task. Based on this observation, we propose to classify the distraction modes, as represented in this visual and cognitive space, to characterize the driver behaviors. Instead of attempting to identify different secondary tasks, which is the focus of most of the driver distraction studies [6], [26], [31], [37], recognizing the proposed distraction modes provides a more general and meaningful representation that can be easily estimated every time a new task or in-vehicle device is introduced. This new representation would also merge different secondary tasks that induce similar distraction level, which simplify the representation.

The proposed classification scheme uses the multimodal features listed on Table II to recognize the distraction modes of the 10 sec. videos. This four-class machine learning problem is implemented using labels derived from the cognitive and visual scores of the videos and the clusters defined in Figure 10. The analysis only includes the 395 videos in which CERT can at least process 66% of the frames (see discussion in Sec. V-A), which increases the reliability of extracted features. The evaluation uses the same six classifiers and follows the same approach to define the training, development and testing partitions described in Section V-C. We optimize the number of features using the development set across folds. Due to the unbalanced mode distribution (see Fig. 10), we report the average precision, average recall and F-score across folds.

Table VII gives the classification results for the five classifiers. It also lists the mean and standard deviation of the number of features used across folds. The average precision and recall rate indicates the complexity of the machine learning problem. Notice that the clusters are defined with data-driven methods. Therefore, it is expected that samples in the boundaries of the clusters may be misclassified. However, the F-scores of the classifiers are significantly higher than chances (i.e., 25%), so we conclude that the features provide discriminative information about the distraction modes.

TABLE VII
DISTRACTION MODE RECOGNITION FOR DIFFERENT CLASSIFIERS (FOUR CLASSES). RESULTS ARE GIVEN IN TERMS OF AVERAGE PRECISION (**P**), AVERAGE RECALL (**R**) AND F-SCORE (**F**).

	Feat #		**P**	**R**	**F**
	Mean	**Std**	[%]	[%]	
LDC	14.3	1.9	51.4	51.2	0.513
KNN	6.1	1.3	45.5	41.6	0.435
SVM1	18.4	1.5	40.4	47.2	0.435
SVM2	11.1	2.1	42.1	42.0	0.421
QDC	11.8	1.7	47.8	48.6	0.482
RUSB	-	-	54.9	54.9	0.549

TABLE VIII
CONFUSION MATRIX OF RUSB FOR DISTRACTION MODE. THE TABLE LISTS THE ACCUMULATED RESULTS ACROSS THE 20 FOLDS. IT ALSO GIVES THE NUMBER OF VALID SAMPLES PER CLUSTER.

			Predicted			
		# Samples	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Actual	Cluster 1	130	**95**	9	21	5
	Cluster 2	90	18	**32**	17	23
	Cluster 3	86	20	11	**50**	5
	Cluster 4	89	7	21	14	**47**

RUSB provides the best performance among the six classifiers with an average F-score equal to 0.549 across the 20 folds (see Table VII). Table VIII presents its confusion matrix to further understand the performance. The matrix is estimated by accumulating the results of the testing partition across the 20 folds (all valid samples are included). Table VIII shows that clusters 1 (LVLC) and cluster 3 (LVMC) are often confused. This is expected since both clusters have similar visual distraction levels. Cluster 2 (MVMC) is adjacent to the other three clusters. There are many samples in the boundaries, especially between clusters 2 and 3, and clusters 2 and 4. These ambiguous samples make the recognition of this cluster the most difficult task, showing the lowest performance (see also Table IX). Despite the mislabeled samples, the numbers on the diagonal of the confusion matrix are the maximum in their respective row and column. This result suggests that the multimodal features provide discriminatory information about the distraction modes over the cognitive and visual space.

Finally, we trained binary classifiers to detect specific distraction modes (i.e., one cluster versus the other three clusters). For real applications, it may be interesting to determine whether the behaviors of a driver belong to one specific distraction mode. For example, an active safety system may need to detect when the driver's behaviors are not longer in cluster 1 (LVLC), which can be used as an indicator of distracted driving behavior. Likewise, it may be relevant to identify when the driver's behaviors are on cluster 4 (HVHC), signaling high distracted behaviors. We address these problems with binary classifiers for each of the distraction modes. These binary classifiers are trained using the same machine learning algorithms (LDC, KNN, SVM1, SVM2, QDC, RUSB) and following the same evaluation procedure as before (i.e., data partition, two-layer driver-independent, cross-validation, feature selection). We form four separate problems in which one distraction mode is classified against the other three.

Table IX shows the results of the distraction mode detection. The table only provides the classifier with the best performance per task, as measured by F-score. It also provides the ten most frequently selected features across folds for each binary classification problem. While the classes for each binary problem are unbalanced (i.e., data from three of the clusters are grouped together), the table shows that the classifiers provide reasonable performance. The binary classifiers achieve F-scores of 0.78 and 0.74 for clusters 1 (LVLC) and 4 (HVHC), respectively. The performances are lower for clusters 2 (MVMC) and 3 (LVMC).

The features that are most frequently selected for these binary classification problems are consistent with the results discussed in Sections V-B and V-C. Most of the features provide discriminative information about either visual or cognitive distractions, it it is expected that they are also useful in detecting the distraction modes. We also noticed that features related to audio and lip movements are often chosen for the detection of cluster 3 (LVMC). This finding is expected since most of the secondary tasks in cluster 3 are associated with *Phone - Talking* and *Conversation* (see Fig. 11(c)).

VII. DISCUSSION AND CONCLUSIONS

This study addressed the problem of driver distraction in terms of both visual and cognitive distractions. Based on real-world driving data, we conducted subjective evaluation from external observers to separately assess both drivers' cognitive and visual distractions. These distraction metrics are validated by the high inter-evaluator agreement observed for both distraction dimensions. Using the evaluations as dependent variables, we trained regularized regression models to predict the perceived distraction level of the drivers. The independent variables are features extracted from the CAN-Bus signal, microphone array and two video cameras facing the road and the driver. The analysis revealed the multimodal features that are linearly related to cognitive and visual distractions. These features were also employed to train classifiers that aim to discriminate between low or high distraction levels. Finally, we presented a clustering analysis of the perceptual evaluations, which shows that the evaluations can be grouped into four different distractions modes. This joint cognitive-visual representation describes different levels of visual and cognitive distractions. We evaluated the performance of multi-class and binary classification problems to recognize these distraction modes, achieving promising results. This novel cognitive and visual representation and the automatic classification of driving behaviors into the proposed distraction modes offer an alternative paradigm to evaluate the detrimental effects caused by different secondary tasks. These tools are especially useful to evaluate new in-vehicle technologies.

The study opens new opportunities in the field of monitoring driving behaviors. An interesting research question is the generalization of the models with new secondary tasks not included in the training of the models. We expect that the regression models will be able to predict the perceived distraction scores even in the presence of new tasks inducing different cognitive or visual workload. Likewise, we can augment the

TABLE IX
DISTRACTION MODE DETECTION USING BINARY CLASSIFICATION (I.E., ONE CLUSTER AGAINST THE OTHERS). THE RESULTS ARE GIVEN IN TERMS OF AVERAGE PRECISION (**P**), AVERAGE RECALL (**R**) AND F-SCORE (**F**). WE REPORT THE CLASSIFIER WITH THE BEST PERFORMANCE.

	Best	**P**	**R**	**F**	**Ten most frequently selected features across folds**
	Classifier	[%]	[%]		
Cluster 1 (LVLC)	RUSB	77.6	78.4	0.780	Yaw Mean; Jaw Drop (AU26) Max; Dimpler (AU14) Min; Lip Tightener (AU23) IQR; Outer Brow Raiser (AU2) Max; Outer Brow Raiser (AU2) Mean; Lip Puckerer (AU18) Max; EOR Dur.; Road Intensity STD; Blink (AU45) Mean
Cluster 2 (MVMC)	RUSB	60.0	63.0	0.615	Blink (AU45) Kurtosis; Yaw IQR; Jitter Kurtosis; Lip Corner Depressor (AU15) Max; Jitter Mean; Lid Tightener (AU7) Min; Road Optical Flow Mean; Road Optical Flow Max; Lip Corner Puller (AU12) Max; Speed Kurtosis
Cluster 3 (LVMC)	RUSB	66.8	70.2	0.684	Cheek Raiser (AU6) IQR; Lip Tightener (AU23) IQR; Brake IQR; Audio GT6 Mean; Audio GT7 Min; Lip Corner Depressor (AU15) Range; Chin Raiser (AU17) IQR; Roll Mean; Lip stretcher (AU20) Skewness; Lip Corner Depressor (AU15) STD
Cluster 4 (HVHC)	LDC	77.0	71.3	0.740	Outer Brow Raiser (AU2) Range; Roll Kurtosis; Lip Tightener (AU23) IQR; Blink Freq.; EOR Freq.; Jaw Drop (AU26) Max; Blink (AU45) Kurtosis; Lips part (AU25) Range; Brake Press Sensor Min; Lip Puckerer (AU18) Kurtosis

proposed representation with other types of distractions (i.e., manual and auditory). The proposed approach can be used to assess in-vehicle technologies where the safety criteria is measured in the multi-dimension distraction space. From an application perspective, an interesting open question is how to define the number of distraction modes that are suitable to represent distractions. The proposed distraction modes are automatically derived from perceptual evaluations. These data-driven clusters can be refined with insights from human factor studies to identify distraction modes that are either acceptable or detrimental to driving safety. In addition, the features identified by this study can be incorporated in the *advanced driver assistance system* (ADAS) to detect distracted driving behaviors. We can also incorporate new features such as fuel consumption and contextual information. These features have to be robustly extracted under different traffic congestions, illuminations, and types of road. More naturalistic driving recordings are needed to address these problems. These research directions can have a positive impact on in-vehicle safety systems that continuously monitor the driver behaviors with noninvasive sensors.

ACKNOWLEDGMENT

The authors thank Dr. John Hansen for his support with the UTDrive Platform. We want to thank the Machine Perception Lab at UCSD for providing the CERT software.

REFERENCES

[1] NHTSA, "Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices," National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT), Washington, D.C., USA, Technical Report NHTSA-2010-0053, April 2013.

[2] M. Kutila, M. Jokela, G. Markkula, and M. Rue, "Driver distraction detection with a camera vision system," in *IEEE International Conference on Image Processing (ICIP 2007)*, vol. 6, San Antonio, Texas, USA, September 2007, pp. 201–204.

[3] Y. Liang, M. Reyes, and J. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 340–350, June 2007.

[4] F. Tango and M. Botta, "Evaluation of distraction in a driver-vehicle-environment framework: An application of different data-mining techniques," in *Advances in Data Mining. Applications and Theoretical Aspects*, ser. Lecture Notes in Computer Science, P. Perner, Ed. Berlin, Germany: Springer Berlin / Heidelberg, 2009, vol. 5633, pp. 176–190.

[5] T. Ersal, H. Fuller, O. Tsimhoni, J. Stein, and H. Fathy, "Model-based analysis and classification of driver distraction under secondary tasks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 692–701, September 2010.

[6] J. Jain and C. Busso, "Analysis of driver behaviors during common tasks using frontal video camera and CAN-Bus information," in *IEEE International Conference on Multimedia and Expo (ICME 2011)*, Barcelona, Spain, July 2011.

[7] A. Azman, Q. Meng, and E. Edirisinghe, "Non intrusive physiological measurement for driver cognitive distraction detection: Eye and mouth movements," in *International Conference on Advanced Computer Theory and Engineering (ICACTE 2010)*, vol. 3, Chengdu, China, August 2010.

[8] Y. Zhang, Y. Owechko, and J. Zhang, "Driver cognitive workload estimation: A data-driven perspective," in *IEEE Intelligent Transportation Systems*, Washington, D.C., USA, October 2004, pp. 642–647.

[9] T. Ranney, W. Garrott, and M. Goodman, "NHTSA driver distraction research: Past, present, and future," National Highway Traffic Safety Administration, Technical Report Paper No. 2001-06-0177, June 2001.

[10] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. Hansen, "UT-Drive: Driver behavior and speech interactive systems for in-vehicle environments," in *IEEE Intelligent Vehicles Symposium*, Istanbul, Turkey, June 2007, pp. 566–569.

[11] Y. Peng, L. Boyle, and S. Hallmark, "Driver's lane keeping ability with eyes off road: Insights from a naturalistic study," *Accident Analysis & Prevention*, vol. 50, pp. 628–634, January 2013.

[12] D. Strayer, J. Watson, and F. Drews, "Cognitive distraction while multitasking in the automobile," in *The Psychology of Learning and Motivation*, B. Ross, Ed. Burlington, MA, USA: Academic Press, February 2011, vol. 54, pp. 29–58.

[13] Q. Wu, "An overview of driving distraction measure methods," in *IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design (CAID CD 2009)*, Wenzhou, China, November 2009.

[14] K. Young, M. Regan, and J. Lee, "Measuring the effects of driver distraction: Direct driving performance methods and measures," in *Driver distraction: theory, effects, and mitigation*, M. Regan, J. Lee, and K. Young, Eds. Boca Raton, FL, USA: CRC Press, October 2008, pp. 85–105.

[15] J. Engström, E. Johansson, and J. Östlund, "Effects of visual and cognitive load in real and simulated motorway driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 2, pp. 97 – 120, March 2005.

[16] K. M. Bach, M. Jaeger, M. Skov, and N. Thomassen, "Interacting with in-vehicle systems: understanding, measuring, and evaluating attention," in *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, Cambridge, United Kingdom, September 2009.

[17] W. Verwey and H. Veltman, "Detecting short periods of elevated workload. a comparison of nine workload assessment techniques," *Journal of Experimental Psychology: Applied*, vol. 2, no. 3, pp. 270–285, September 1996.

[18] J. Jain and C. Busso, "Assessment of driver's distraction using perceptual evaluations, self assessments and multimodal feature analysis," in *5th Biennial Workshop on DSP for In-Vehicle Systems*, Kiel, Germany, September 2011.

[19] N. Li and C. Busso, "Using perceptual evaluation to quantify cognitive and visual driver distractions," in *Smart Mobile In-Vehicle Systems – Next Generation Advancements*, G. Schmidt, H. Abut, K. Takeda, and J. H. L. Hansen, Eds. New York, NY, USA: Springer, January 2014, pp. 183–207.

[20] W. Piechulla, C. Mayser, H. Gehrke, and W. König, "Reducing drivers' mental workload by means of an adaptive man–machine interface,"

Transportation Research Part F: Traffic Psychology and Behaviour, vol. 6, no. 4, pp. 233–248, December 2003.

[21] C. Busso and J. Jain, “Advances in multimodal tracking of driver distraction,” in *Digital Signal Processing for In-Vehicle Systems and Safety*, J. Hansen, P. Boyraz, K. Takeda, and H. Abut, Eds. New York, NY, USA: Springer, December 2011, pp. 253–270.

[22] A. Sathyanarayana, S. Nageswaren, H. Ghasemzadeh, R. Jafari, and J. Hansen, “Body sensor networks for driver distraction identification,” in *IEEE International Conference on Vehicular Electronics and Safety (ICVES 2008)*, Columbus, OH, USA, September 2008.

[23] D. Chiang, A. Brooks, and D. Weir, “On the highway measures of driver glance behavior with an example automobile navigation system,” *Applied Ergonomics*, vol. 35, no. 3, pp. 215–223, May 2004.

[24] C. Patten, A. Kircher, J. Östlund, L. Nilsson, and O. Svenson, “Driver experience and cognitive workload in different traffic environments,” *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 887–894, September 2006.

[25] T. Lansdown, N. Brook-Carter, and T. Kersloot, “Distraction from multiple in-vehicle secondary tasks: vehicle performance and mental workload implications,” *Ergonomics*, vol. 47, no. 1, pp. 91–104, January 2004.

[26] H. Alm and L. Nilsson, “The effects of a mobile telephone task on driver behaviour in a car following situation,” *Accident Analysis & Prevention*, vol. 27, no. 5, pp. 707–715, October 1995.

[27] W. Tseng, H. Nguyen, J. Liebowitz, and W. Agresti, “Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files,” *Industrial Management & Data Systems*, vol. 105, no. 9, pp. 1188–1205, 2005.

[28] F. Tango, L. Minin, F. Tesauri, and R. Montanari, “Field tests and machine learning approaches for refining algorithms and correlations of driver’s model parameters,” *Applied Ergonomics*, vol. 41, no. 2, pp. 211–224, March 2010.

[29] M. Miyaji, M. Danno, H. Kawanaka, and K. Oguri, “Driver’s cognitive distraction detection using adaboost on pattern recognition basis,” in *IEEE International Conference on Vehicular Electronics and Safety (ICVES 2008)*, Columbus, OH, USA, September 2008, pp. 51–56.

[30] J. Lee, M. Reyes, T. Smyser, Y. Liang, and K. Thornburg, “SAfety VEhicles using adaptive interface technology (task 5) final report: Phase 1,” The University of Iowa, Iowa City, IA, USA, Technical Report, November 2004.

[31] N. Li, J. Jain, and C. Busso, “Modeling of driver behavior in real world scenarios using multiple noninvasive sensors,” *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1213–1225, August 2013.

[32] N. Li and C. Busso, “Analysis of facial features of drivers under cognitive and visual distractions,” in *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, July 2013.

[33] P. Angkititrakul, D. Kwak, S. Choi, J. Kim, A. Phucphan, A. Sathyanarayana, and J. Hansen, “Getting start with UTDrive: Driver-behavior modeling and assessment of distraction for in-vehicle speech systems,” in *Interspeech 2007*, Antwerp, Belgium, August 2007, pp. 1334–1337.

[34] V. Neale, T. Dingus, S. Klauer, J. Sudweeks, and M. Goodman, “An overview of the 100-car naturalistic study and findings,” National Highway Traffic Safety Administration, Technical Report Paper No. 05-0400, June 2005.

[35] A. Sathyanarayana, P. Boyraz, Z. Purohit, R. Lubag, and J. Hansen, “Driver adaptive and context aware active safety systems using CAN-bus signals,” in *IEEE Intelligent Vehicles Symposium (IV 2010)*, San Diego, CA, USA, June 2010.

[36] M. Slaney, “An efficient implementation of the patterson-holdsworth auditory filter bank,” Apple Computer, Perception Group, Tech. Rep. 35, 1993.

[37] T. Horberry, J. Anderson, M. Regan, T. Triggs, and J. Brown, “Driver distraction: the effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance,” *Accident Analysis & Prevention*, vol. 38, no. 1, pp. 185–191, January 2006.

[38] K. A. Brookhuis and D. de Waard, “Assessment of drivers’ workload: Performance and subjective and physiological indexes,” in *Stress, workload, and fatigue*, ser. Human Factors in Transportation, P. Hancock and P. Desmond, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates Inc., November 2000, pp. 321–333.

[39] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Automatic recognition of facial actions in spontaneous expressions,” *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, September 2006.

[40] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, “Drowsy driver detection through facial movement analysis,” in *Human-Computer Interaction*, ser. Lecture Notes in Computer Science, M. Lew, N. Sebe, T. Huang, and E. Bakker, Eds. Berlin, Germany: Springer Berlin / Heidelberg, December 2007, vol. 4796, pp. 6–18.

[41] J. Whitehill and J. Movellan, “A discriminative approach to frame-by-frame head pose tracking,” in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008)*, Amsterdam, The Netherlands, September 2008.

[42] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[43] J. McCall and M. Trivedi, “Driver behavior and situation aware brake assistance for intelligent vehicles,” *Proceedings of the IEEE*, vol. 95, no. 2, pp. 374–387, February 2007.

[44] P. Caffier, U. Erdmann, and P. Ullsperger, “Experimental evaluation of eye-blink parameters as a drowsiness measure,” *European Journal of Applied Physiology*, vol. 89, no. 3-4, pp. 319–325, May 2003.

[45] C. VanVoorhis and B. Morgan, “Understanding power and rules of thumb for determining sample sizes,” *Tutorials in Quantitative Methods for Psychology*, vol. 3, no. 2, pp. 43–50, 2007.

[46] N. Li and C. Busso, “Driver mirror-checking action detection using multi-modal signals,” in *The 6th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems*, Seoul, Korea, September-October 2013, pp. 101–108.

[47] C. Seiffert, T. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “RUSBoost: Improving classification performance when training data is skewed,” in *International Conference on Pattern Recognition (ICPR 2008)*, Tampa, FL, USA, December 2008.

[48] E. Hruschka, R. Campello, A. Freitas, and A. De Carvalho, “A survey of evolutionary algorithms for clustering,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 39, no. 2, pp. 133–155, March 2009.

Nanxiang Li (S’2012) received his B.S. degree (2005) in Electrical Engineering from Xiamen University, Fujian, China. He received his M.S. degree (2009) in Electrical Engineering from University of Alabama, Tuscaloosa, Alabama, USA. He is currently pursuing the Ph.D. degree in Electrical Engineering at the University of Texas at Dallas (UTD), Richardson, Texas, USA. He joined the Multimodal Signal Processing (MSP) laboratory at UTD in 2011. His research interests include human attention modeling in the context of driving behavior analysis, and multimodal interfaces with emphasis on gaze estimation. He has also worked on human tracking and recognition using gait information.

Carlos Busso (S’02-M’09, SM’13) is an Assistant Professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He received his B.S (2000) and M.S (2003) degrees with high honors in electrical engineering from University of Chile, Santiago, Chile, and his Ph.D (2008) in electrical engineering from University of Southern California (USC), Los Angeles, USA. He was selected by the School of Engineering of Chile as the best Electrical Engineer graduated in 2003 across Chilean universities. At USC, he received a Provost Doctoral Fellowship from 2003 to 2005 and a Fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests are in digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of in-vehicle active safety system, affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, and machine learning methods for multimodal processing.