IEEE 18th International Conference on Intelligent Transportation Systems, 2015 - To Appear

Automatic Critical Event Extraction and Semantic Interpretation by Looking-Inside

Sujitha Martin, Eshed Ohn-Bar and Mohan M. Trivedi

***Abstract*—Data-driven systems are becoming prevalent for driver assistance systems and with large scale data collection such as from the 100-Car Study and the second Strategic Highway Research Program (SHRP2), there is a need for automatic extraction of critical driving events and semantic characterization of driving. This is especially necessary for videos looking at the driver since manual extraction and annotation is time-consuming and subjective. This labeling process is often overlooked and undervalued, even though data mining is the first critical step in the design and development of machine vision based algorithms for predictive, safety systems. In this paper, we define and implement quantitative measures of vocabularies often used by data reductionist when labeling videos of looking at the driver. This is demonstrated on a significantly large sum of data containing almost 200 minutes (600 000 frames total from two videos of looking-in) of multiple drivers collected by UCSD-LISA. We qualitatively show the advantages of automatically extracting such information on this relatively large scale data.**

I. INTRODUCTION

Data on driving come in many forms, including police reports on crashes, driving simulators, controlled on-road driving and naturalistic on-road driving. In recent years, huge efforts have been taken to collect naturalistic driving data such as the 100-car study [1] and the Second Strategic Highway Research Program (SHRP2) [2]. Naturalisitic driving studies (NDS) datasets are generally designed to capture natural driving of everyday drivers using non-intrusive sensors (e.g. cameras). There is a need for such data in order to understand the causes of crashes and near-crashes [3], and to develop advanced predictive and active safety driver assistance systems to prevent or mitigate traffic accidents.

In order to study large-scale naturalistic driving data such as SHRP2 and 100-Car, there is a need for automatic reduction of video data, which is by nature a more complex representation of the state of the driver than, say, vehicle dynamics parameters. Currently, human reductionists are following guidelines similar to the ones presented in [4] for annotating precipitating events leading to crashes, near-crashes or incidents. For example, one category of annotation is called "Inattention to forward roadway". For example, lets assume it takes a human reductionist 10 seconds per video frame to annotate where the driver is looking (Sivaraman and Trivedi [5] reported that it took 27.8 hours to annotate vehicle locations in 10 000 video frames which translates to 10 seconds of annotation per video frame). A one hour video captured at 25 frames per second contains 9 000 000 frames. This translates to 25 000 annotation hours with no breaks. This is only considering one hour of video compared to the 43 000 hours of data in the 100-car study and more than 1 000 000 hours of data in SHRP2. Furthermore, labeling where the driver is looking is only a small part of the data reductionist dictionary. Table I lists a few out of the at least 100 categories defined in [4] with respect to looking-inside.

Manually annotating all parts of big data is humanly impossible in a time frame necessary to design and deploy safety measures. Therefore, there is a need for **automating the low-level process of labeling data** and **deriving mid-level semantic characteristics**, such as this recent publication [6] on data reduction by looking at lane information and vehicle dynamics. Looking-outside has been studied for decades and the dictionary/vocabulary to extract key events are well defined. On the other hand, looking-inside is still relatively new and more importantly still lacks the concrete definition for data reduction. In this paper, we adopt a few event definitions from the SHRP2 data dictionary [4] such as inattention to forward roadway and number of hands on the wheel. We then extract these events from large scale driving data and generate semantic driving reports (i.e. number of events, the duration of events, frequency of events). These automatically extracted events will improve the efficiency in data mining, which is necessary when developing safety critical systems, and the driving reports will be useful for driver/driving style recognition and quality analysis.

TABLE I. LIST OF SELECTED ELEMENTS IN DATA DICTIONARY [4] FROM LOOKING-INSIDE FOR USE IN NDS

Category
Drowsy, sleepy, asleep, fatigued
Talking/singing
Reaching for object
Texting on cell phone
Adjusting/monitoring devices integral in instrument cluster (e.g.radio)
Inattention to forward roadway (e.g. left window)
Inattention to forward roadway (e.g. right window)
No. of hands on the wheel (i.e. no hands, 1 hand, 2 hands)
Use of seat belt (e.g. lap only)

II. RELATED WORKS

In a 2006 report on the results of 100-car field experiment [7], it showed that almost 80 percent of all crashes and 65 percent of all near-crashes involved driver looking away from the forward roadway just prior to the incident. It also revealed that 67% of crashes and 82% of near-crashed occurred when subjects were driving with at least one hand on the wheel. These deductions resulted from a data reduction process were employees inspect segments of video related to crashes and near crashes, and label the event with predefined classes that provides information regarding the causes leading up to the conflict. Manual data reduction of data is time-consuming and is subject to interpretation of the reductionist. Automating this process, especially video processing [8], will allow for analysis

Authors are with the Laboratory of Intelligent and Safe Automobiles (http://cvrr.ucsd.edu/members/), UC San Diego, La Jolla, CA, USA

of large scale data in a reasonable time frame necessary for developing predictive safety systems.

Recently, there has been a hype in data reduction using vehicle dynamics and looking outside on large scale NDS data. In [6], the authors propose using vehicle dynamics and vision based lane semantics to extract semantic information about vehicle localization within lanes, traffic density and road curvature and more recently, analyze the quality of a subset of the extracted events in [9]. In [10], the authors experiment with event recognition algorithms which automatically extract and label certain patterns in the CAN- and GPS-data. In [11], the authors used forward looking radar and vehicle dynamics to identify events where drivers applied brakes while following another vehicle. However, these studies are purely based on looking outside and vehicle dynamics. To the best of our knowledge, no work exists in literature for automatic reduction of data based on looking-inside at the driver.

Our work is first of the kind in using looking in at the driver as cues to extract key events from large scale driving data and deriving semantic driving characteristics. Extracting events based on looking at the driver is a very useful tool in the development of driver assistance systems because it expedites the time it takes to gather such events to learn predictive models. For example, automatic extraction of safety critical events based on looking-inside will be very useful in the development of head-and-hand-coordinated cues for driver activity classification [12] and early prediction of maneuvers such as overtake and break events [13].

Similarly deriving semantic driving characteristics from a drive is useful for driver or driving style recognition. There are many works in literature where driving style recognition using vehicle dynamics. Johnson and Trivedi [14] used cues from vehicle mounted smartphone sensors to characterize driving style into one of two categories: “normal” (non-aggressive) and aggressive. Ly et al. [15] showed that input from vehicle’s inertial sensors can be used to build a profile of the driver to eventually give feedback to the driver. These and many other works attempt to recognize driving style based on vehicle dynamics. In this work, we show that driving style can also be defined by looking at the driver, including the way a driver observes the surround (e.g. number of head scans) and how many hands are on the wheel (e.g. is the driver mostly using one hand).

III. Small Modules for Big Data Analysis

A. Low-level Feature Extraction

Low-level features of interest to extract from the videos of looking inside the vehicle cockpit include driver’s surrogate gaze direction (i.e. head pose) and the number of driver’s hands on the wheel. The contribution in this paper is to define and show by example the potential in automatic labeling of low-level features and deriving mid-level semantic information on large-scale driving data. Algorithms described below are included for the paper to be self-contained.

1) Head Dynamics Analysis: Head pose is estimated by the relative configuration of selective facial landmarks and their corresponding points on a 3D generic face model [16], [17]. The key component is facial landmark estimation and tracking. There are many number of geometric based facial landmark detection methods including constrained local models [18], mixture of pictorial structures [19] and cascaded regression models [20], [21]. In this work, we use the cascade of regression models for facial landmark estimation. The idea is given an initial estimate of the facial landmark locations, say p_0, which is generally a mean shape, and a learned sequence of regression models $(R_1, ..., R_K)$, facial landmark location at the kth iteration is computed as follows:

$$p_k = p_{k-1} + R_k * F(p_{k-1}, M)$$

where $k \in [1, K]$, K is the maximum number of iterations, M represents the image on which the landmarks are being estimated, p_{k-1} is the landmark positions at $(k-1)$ iteration and $F(p_{k-1}, I)$ is a vector of features extracted at locations $p_(k-1)$ on image M. From the detected facial landmarks, selected points (i.e. eye corners, nose corners and nose tip) and their 3D points on a generic face model are used as inputs to the pose from orthography and scaling (POS) [22] algorithm to estimate the head pose, as shown in Fig. 1.

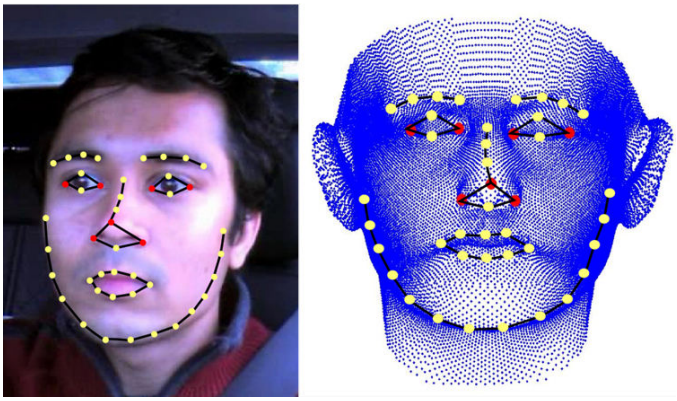


Fig. 1. Illustration of the relative facial landmarks on a 2D image and 3D generic face model. The red points are key facial landmarks used along with the POS algorithm to compute head pose.

The estimated head pose orientation is with respect to the camera coordinate system. Since the driver face viewing camera is biased to the right of the driver, when the driver is facing forward the estimated yaw rotation angle is closer to 20 degrees. Therefore, in order to remove this bias, if $R(t)$ defines the biased head rotation matrix at time t and R_0 represents the head rotation when the driver is facing forward, then the unbiased head rotation matrix is given by:

$$R^*(t) = R_0^T * R(t)$$

where $R_0^T * R_0 = I$, the identity matrix.

2) Hand Activity Analysis: For hand activity, we follow the method proposed in [23], outlined in Fig. 2. First, a set of image regions are defined which may be of interest for studying driver actions. In this paper, these are three rectangular regions around the wheel, gear shift, and instrument panel. The method was shown beneficial over motion-based hand tracking in [24], and leverages cues from multiple regions in the scene in order to resolve visually challenging settings. Each region is represented using shape cues, as following. The horizontal and vertical edge image of each region is split into a uniformly spaced spatial grid, and a histogram of edges by

orientation is performed for each grid cell. The histograms are L2-normalized and concatenated. This process repeats with varying grid sizes in order to capture details of the hand at multiple spatial resolutions. The visual descriptors for each region of interest are concatenated in an early fusion manner, and a multiples Support Vector Machine [25] is used to map the visual descriptors to a hand activity class in the activity vocabulary regarding the number of hands in each region.

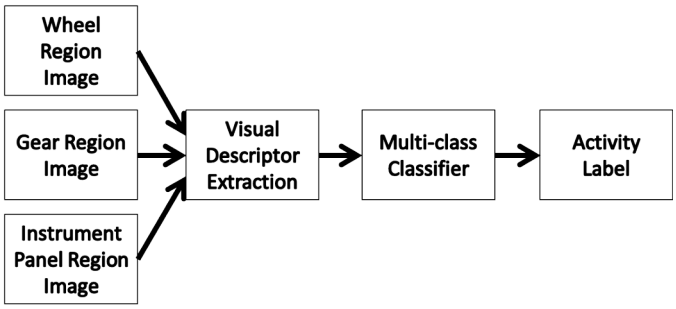


Fig. 2. Early-fusion scheme where a multiclass classifier receives the visual features and performs the activity classification

B. Mid-level Semantic Extraction

Following are formal definitions of semantic understanding derived from using the above described low-level features on frame sequences of looking-inside the vehicle. First lets define the variables that will be used:

*1) **Variables**:*

- N = total number of frames in the video sequence
- n = nth frame in the video, $1 \leq n \leq N$
- $head(n)$ is head pose in the yaw rotation angle at instance n where a positive value is leftward facing and a negative value is rightward facing
- τ_{right} = right threshold
- τ_{left} = left threshold
- $F(n)$ is the facing direction of the driver at instance n with three possible states $\{\text{Left}, \text{Forward}, \text{Right}\}$

*2) **State definitions**:*

- Facing state for driver:

$$F(n) = \begin{cases} \text{Left} & \tau_{left} < head(n) \\ \text{Forward} & \tau_{right} \leq head(n) \leq \tau_{left} \\ \text{Right} & head(n) < \tau_{right} \end{cases}$$

- $hands_W(n)$ represents the number of hands on the wheel, with three possible values $\{0, 1, 2\}$

*3) **Characteristic definitions**:* For a given video sequence

- *Facing forward (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{F(n) == \text{Forward}\}$$

- *Facing left (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{F(n) == \text{Left}\}$$

- *Facing right (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{F(n) == \text{Right}\}$$

- *No. of left head turn events*: Number of continuous segments where the driver state ($L(n)$) is "Left".
- *No. of right head turn events*: Number of continuous segments where the driver state ($L(n)$) is "Right".
- *Mean duration of facing left (or right) (in milliseconds)*: Let duration of a continuous sequence of facing left (or right) be such that during the segment it's always left (or right) looking . The mean is then taken over duration of all continuous sequences of facing left (or right) labels.
- *Frequency of facing left (or right) (per minute)*: The number of left (or right) head turn events divided by the total recording time of the video sequence in minutes.
- *Two hands on the wheel (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{hands_W(n) == 2\}$$

- *One hand on the wheel (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{hands_W(n) == 1\}$$

- *No hands on the wheel (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{hands_W(n) == 0\}$$

- *No. of two hands on the wheel events*: Number of continuous segments where the driver has two hands on the wheel.
- *No. of one hand on the wheel events*: Number of continuous segments where the driver has one hand on the wheel.
- *No. of no hands on the wheel events*: Number of continuous segments where the driver has no hands on the wheel.
- *Mean duration of two hands (or one hand or no hands) on the wheel (seconds)*: Let duration of a continuous sequence of two hands (or one hand or no hands) on the wheel be such that during the segment it's always two hands on the wheel . The mean is then taken over duration of all continuous sequences of two hands (or one hand or no hands) on the wheel labels.
- *Frequency of two hands, one hand or no hands on the wheel (per minute)*: The number of two hands, one hand or no hands, respectively, on the wheel

events divided by the total recording time of the video sequence in minutes.

- *Two hands on the wheel and forward facing (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{hands_W(n) == 2\} * \mathbb{1}\{F(n) == \text{Forward}\}$$

- *One hand on the wheel and forward facing (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{hands_W(n) == 1\} * \mathbb{1}\{F(n) == \text{Forward}\}$$

- *One hand on the wheel and not forward facing (%)*:

$$\frac{1}{N} * \sum_{n=1}^{N} \mathbb{1}\{hands_W(n) == 1\} * \mathbb{1}\{F(n)! = \text{Forward}\}$$

IV. Drive Analysis on Naturalistic Driving Data

A. Dataset Collection and Statistics

The data used in this work comes from a vehicle instrumented at the UCSD-LISA laboratory called the LISA-Q2. LISA-Q2 is instrumented with 5 camera sensors looking at the driver's face, the driver's hand, the cabin view, the front view and the rear view, the inertial motion unit (IMU) and the global positioning system (GPS). The camera perspectives are intentionally designed to be similar to that which is used in the 100-Car and SHRP2 naturalistic driving studies. In addition, data is also captured from in-vehicle CAN bus. This paper focuses on the analysis of looking-inside and therefore, the data from 2-camera sensors (i.e. driver's face view and driver's hand view) will be analyzed here.

LISA-Q2 has been used for the past two years to collect data in Southern California, primarily in the urban streets and highways surrounding the University of California, San Diego campus. In these past two years, the data collection has amounted to containing at least 14 different drivers of varying age and driving experience, at least 100 trips of more than 20 minutes of driving each, and at least 300 GB of data total (samples along with vision challenges are presented in [26]). It is important to note that all forms of data within a trip are time-synchronized to the millisecond.

In this paper, seven trips from the LISA-Q2 database are used to demonstrate the benefits of automatically labeling low-level features and deriving mid-level semantic information using the modules described earlier. These seven trips are focused on two drivers driving on highways on different days and different times of the day for at least 15 minutes. The subjects are given no instructions on how to drive. More details about the trips, such as the duration of the trip, the number of frames from the head camera sensor and the miles driven, are summarized in II.

TABLE II. Statistics on the eight trips from the LISA-Q2 which are analyzed in this work

Trip No.	Subject	Duration (minutes)	Frames	Miles Driven
1	*A*	16.9	26000	21.5
2	*A*	36.8	56500	58.6
3	*A*	28.1	42750	28.5
4	*A*	30.5	46700	20.7
5	*B*	22.8	35000	21.4
6	*B*	29.7	45300	67.5
7	*B*	25.0	38000	22.9
All Trips		198.6	303750	252.3

B. Drive Analysis: Labeling and Semantic Characteristics

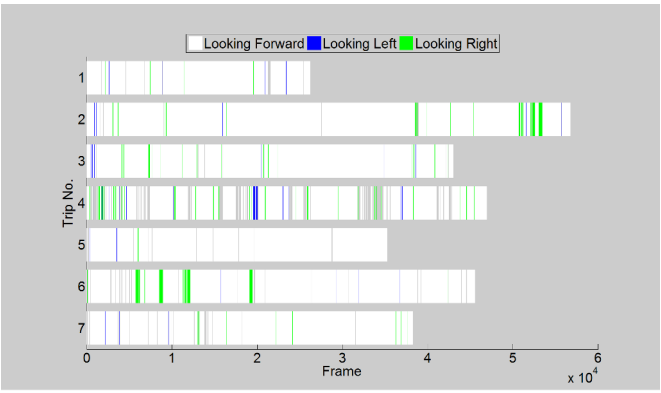
Data can be labeled in many number of ways using low-level features of the driver state, vehicle dynamics and surround information. In this work, we show data labeling using well-known and measurable low-level features of driver head pose and driver hand position. As described in the earlier section, we are analyzing seven trips containing two drivers. The automatic labeling of the seven trips using head cues only, hand cues only and head plus hand clues are shown in Fig. 3. In Fig. 3a is the visualization of labeled data based on features from driver's head analysis, where white represents driver facing forward, blue represents driver facing left and green represents driver facing right. In Fig. 3b is the visualization of labelled data based on features from driver's hand analysis, where blue represents two hands on the wheel, green represents one hand on the wheel and red represents no hands on the wheel. In Fig. 3c is the visualization of labelled data based on features from driver's head and hand analysis, where black represents two hands on the wheel and driver facing forward, yellow represents one hand on the wheel and driver facing forward and magenta represents one hand on the wheel and driver not facing forward (i.e. facing right/left). Note that there are video frames where reliable information was not measurable and thus no color was assigned.

Illustration of labeled data in the manner shown in Fig. 3 serves two purposes. One is to compare different drives and possibly choose which drive to further analyze depending on the presence and frequency of activities. For instance, in illustration of labeled data using head cues alone (Fig. 3a) there is a significant difference between trip no. 2 and no. 4, although they are from the same driver. Second purpose is, within a chosen drive, to determine which part is more interesting or necessary to analyze. For example, given trip number 1, if there is interest in analyzing only where the driver has one hand on the wheel, Fig. 3b depicts where to find those events. Fig. 3c is an interesting illustration which shows when and where the driver is or is not following the general rule of driving of "*Keeping eyes on the road and hands on the wheel*". Fig. 4 shows two sample images taken from automatic labeling of one instance where driver has two hands on wheel and facing forward and another instance where driver has one hand on the wheel and not facing forward. The color coded illustration of labeled data will be very useful when designed in an interactive manner with synchronized video feeds.

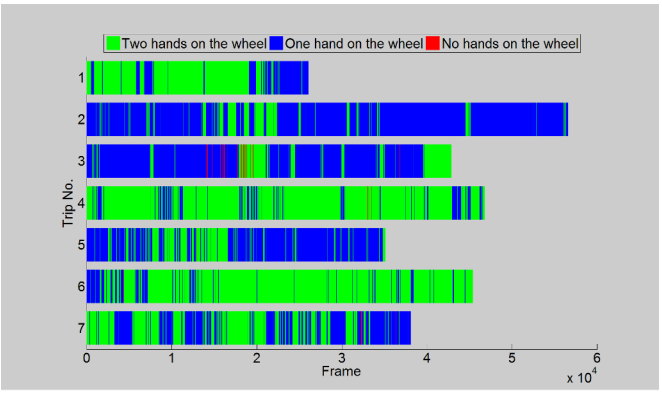
A summary report of drives can also be gathered from the event extraction as shown for selected trips in Table III. This semantic characteristic summary gives a more numeric representation of the automatic labeling of data. This shows how frequently one subject scans the driving environment with respect to others and how was one trip with respect to another. These entries in the summary report can play a key role in driving style recognition by looking-inside.

V. Conclusions

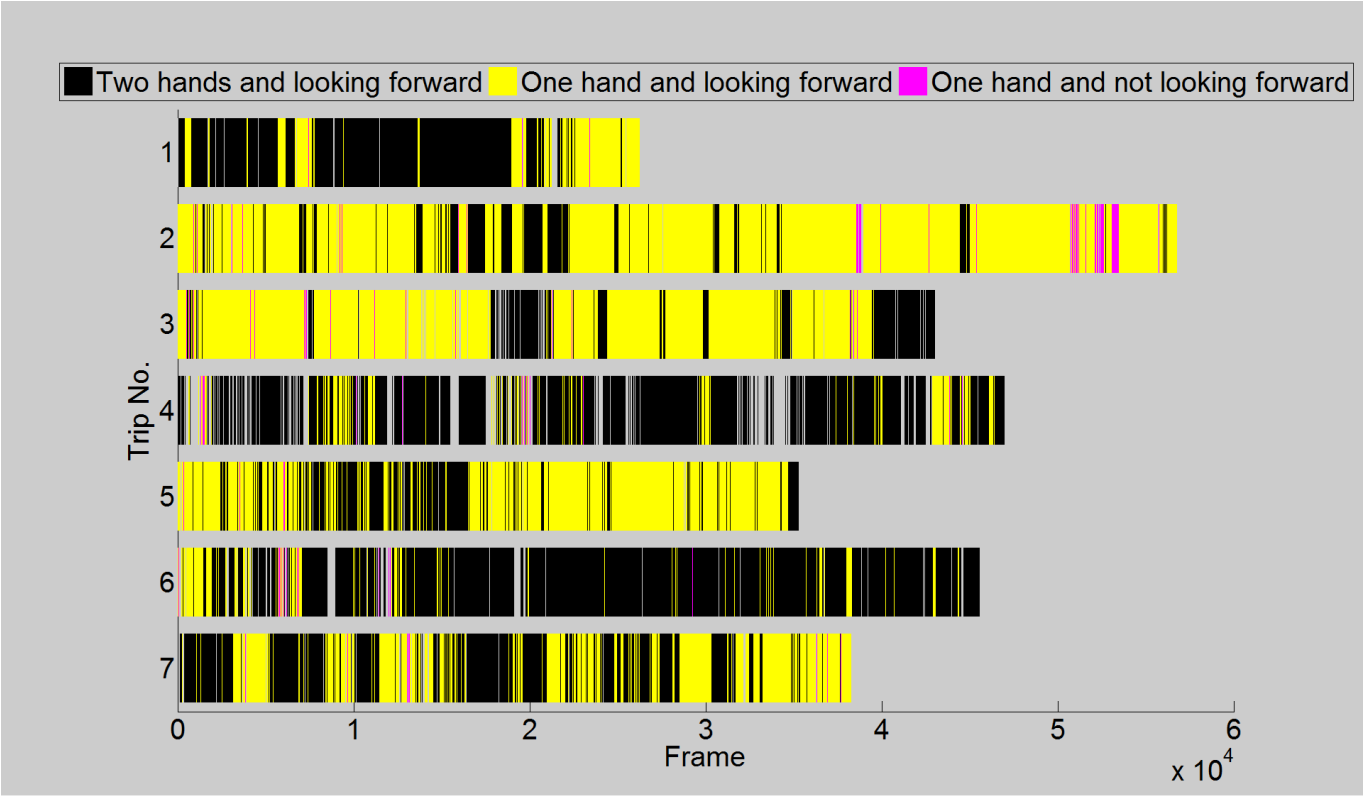
In this work we motivated the need for automatic extraction of critical driving events when looking at the driver inside



(a) Labeled data using driver's head analysis



(b) Labeled data using driver's hand analysis



(c) Labeled data using driver's head and hand analysis

Fig. 3. Visualization of structured data based on features from (a) driver's head analysis, (b) driver's hand analysis and (c) both driver head and hand analysis. Legend: (a) white represents driver facing forward, blue represents driver facing left and green represents driver facing right, (b) blue represents two hands on the wheel, green represents one hand on the wheel and red represents no hands on the wheel, and (c) black represents two hands on the wheel and driver facing forward, yellow represents one hand on the wheel and driver facing forward and magenta represents one hand on the wheel and driver not facing forward (i.e. facing right/left).

the vehicle. Using event definitions as defined in the data reduction dictionary, we defined quantitative measures based on head and hand view cameras. A large scale naturalistic driving data is being collected by the LISA team at the University of California, San Diego, of which approximately 200 minutes, which is 600 000 frames from two videos of looking-inside, is analyzed in this work. The demonstration shows the usefulness of automatic labeling of data and deriving semantic characteristics for driver style recognition.

REFERENCES

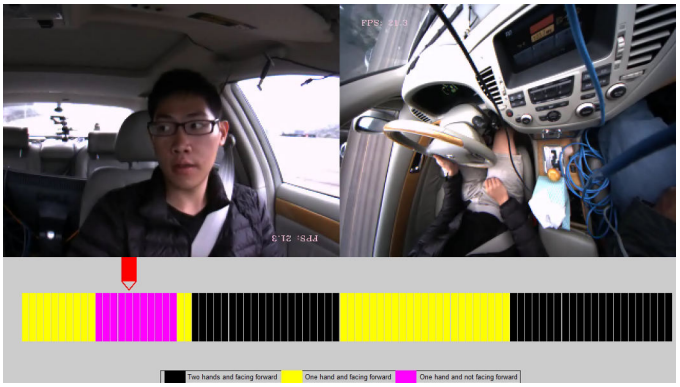
[1] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman, "An overview of the 100-car naturalistic study and findings," Virginia Tech Transportation Institute and National Highway Traffic Safety Administration, Tech. Rep., 2005.

[2] L. N. Boyle, S. Hallmark, J. Lee, D. V. McGehee, D. M. Neyens, and N. J. Ward, "Integration of analysis methods nad development of analysis plan - shrp2 safety research," Transportation Research Board of National Acamdemics, Tech. Rep., 2012.

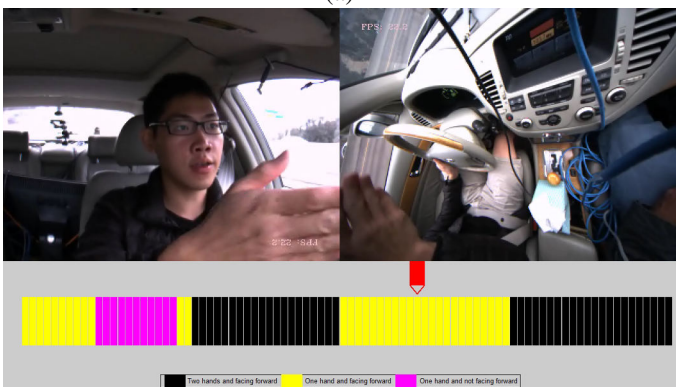
[3] R. Tian, L. Li, M. Chen, Y. Chen, and G. Witt, "Studying the effects of driver distraction and traffic density on the probability of crash and near-crash events in naturalistic driving environment," *Intelligent*

TABLE III. DRIVE ANALYSIS SEMANTICS

Subject	A		B	
Trip No.	1	4	5	7
Facing forward (%)	93	72	97	92
Facing left (%)	1.1	3.9	0.5	2.4
Facing right (%)	2.9	8.6	0.6	2.7
No. of left head turn events	21	102	15	118
No. of right head turn events	41	214	19	55
Mean duration of facing left (milliseconds)	469	659	361	243
Mean duration of facing right (milliseconds)	658	670	403	698
Frequency of facing left (per minute)	1.24	3.33	0.65	4.73
Frequency of facing right (per minute)	2.42	7.00	0.83	2.20
Two hands on the wheel (%)	71	85	31	52
One hand on the wheel (%)	29	15	69	47.9
No hands on the wheel (%)	0	0.2	0.0002	0.004
No. of two hands on the wheel events	114	331	375	303
No. of one hand on the wheel events	114	313	385	299
No. of no hands on the wheel events	0	11	1	6
Mean duration of two hands on the wheel (sec)	6.27	4.65	1.08	2.54
Mean duration of one hand on the wheel (sec)	2.52	0.83	2.42	2.39
Mean duration of no hands on the wheel (sec)	0	0.29	0.19	0.85
Frequency of two hands on the wheel (per min)	6.75	10.84	16.46	12.13
Frequency of one hand on the wheel (per min)	6.75	10.25	16.90	11.97
Frequency of no hands on the wheel (per min)	0	0.36	0.04	0.24
Two hands on the wheel & forward facing (%)	68	61	30	48
One hand on the wheel & forward facing (%)	25	11	67	44
One hand on the wheel & not forward facing (%)	21	2	1	3



(a)



(b)

Fig. 4. Two sampled instances from a sequence of automatically labeled data, where (a) driver has one hand on the wheel and not facing forward and (b) driver has one hand on the wheel and facing forward.

Transportation Systems, IEEE Transactions on, 2013.

[4] "Researcher dictionary for video reduction data," Virginia Tech Transportation Institute, Blacksburg, VA, Tech. Rep., December 2010.

[5] S. Sivaraman and M. M. Trivedi, "Active learning for on-road vehicle detection: A comparative study," *Machine vision and applications*, 2014.

[6] R. K. Satzoda and M. M. Trivedi, "Drive analysis using vehicle dynamics and vision-based lane semantics," *Intelligent Transportation Systems, IEEE Transactions on*, 2014.

[7] T. A. Dingus, S. Klauer, V. Neale, A. Petersen, S. Lee, J. Sudweeks, M. Perez, J. Hankey, D. Ramsey, S. Gupta *et al.*, "The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment," Tech. Rep., 2006.

[8] M. Dozza and N. P. González, "Recognising safety critical events in real traffic: Can automatic video processing be used to improve naturalistic data analyses?" *Accident Analysis & Prevention*, 2013.

[9] R. Satzoda, P. Gunaratne, and M. M. Trivedi, "Drive quality analysis of lane change maneuvers for naturalistic driving studies," in *IEEE Intelligent Vehicles Symposium*, 2015.

[10] M. Benmimoun, F. Fahrenkrog, A. Zlocki, and L. Eckstein, "Incident detection based on vehicle can-data within the large scale field operational test eurofot," in *22nd Enhanced Safety of Vehicles Conference (ESV 2011), Washington, DC/USA*, 2011.

[11] K. D. Kusano, J. Montgomery, and H. C. Gabler, "Methodology for identifying car following events from naturalistic data," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*. IEEE, 2014.

[12] S. Martin, E. Ohn-Bar, A. Tawari, and M. M. Trivedi, "Understanding head and hand activities and coordination in naturalistic driving videos," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*. IEEE, 2014.

[13] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistnace systems," *Computer Vision and Image Understanding*, 2015.

[14] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Intelligent Transportation Systems (ITSC), 14th International IEEE Conference on*. IEEE, 2011.

[15] M. V. Ly, S. Martin, and M. M. Trivedi, "Driver classification and driver style recognition using inertial sensors," in *IEEE Intelligent Vehicles Symposium*, 2013.

[16] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi, "On the design and evaluation of robust head pose for visual user interfaces: algorithms, databases, and comparisons," in *Automotive User Interfaces and Interactive Vehicular Applications*, 2012.

[17] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator (cohmet) for driver assistance: Issues, algorithms and on-road evaluations," *Intelligent Transportation Systems, IEEE Transactions on*, To appear.

[18] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009.

[19] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[20] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013.

[21] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013.

[22] D. F. Dementhon and L. S. Davis, "Model-based object pose in 25 lines of code," *International journal of computer vision*, 1995.

[23] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014.

[24] E. Ohn-Bar, S. Martin, and M. M. Trivedi, "Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies," *Journal of Electronic Imaging*.

[25] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *The Journal of Machine Learning Research*, 2002.

[26] S. Martin, E. Ohn-Bar, A. Moegelmose, K. Yuen, and M. M. Trivedi, "Vision challenges in naturalistic driving videos," CVPR Workshop on The Future of Datasets in Vision, Tech. Rep., 2015.