2014 IEEE Intelligent Vehicles Symposium

(IV 2014)

Ypsilanti, Michigan, USA
8 – 11 June 2014

Pages 719-1468

IEEE Catalog Number: CFP14IVS-POD
ISBN: 978-1-4799-3639-7

2014 IEEE Intelligent Vehicles Symposium

(IV 2014)

Ypsilanti, Michigan, USA
8 – 11 June 2014

Pages 719-1468

IEEE

IEEE Catalog Number: CFP14IVS-POD
ISBN: 978-1-4799-3639-7

******This publication is a representation of what appears in the IEEE Digital Libraries. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number: CFP14IVS-POD
ISBN 13: 978-1-4799-3639-7

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

proceedings.com

TABLE OF CONTENTS

MOORAL SESSION IC1 – IMAGE, RADAR, SIGNAL PROCESSING I

PAPER MOPOSTER SESSION ID1 – AUTOMATED VEHICLES I

MOORAL SESSION IE1 – DRIVER ASSISTANCE SYSTEMS I

TUORAL SESSION IIA1 – AUTOMATED VEHICLES II

TUPOSTER SESSION IIB1 – IMAGE, RADAR, LIDAR SIGNAL PROCESSING II

TUORAL SESSION IIC1 – DRIVER ASSISTANCE SYSTEMS II

WEORAL SESSION IIIA1 – VEHICLE ENVIRONMENT PERCEPTION II

WEPOSTER SESSION IIIB1 – VEHICLE ENVIRONMENT PERCEPTION III

Predicting Driver Maneuvers by Learning Holistic Features

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

Abstract— **In this work, we propose a framework for the recognition and prediction of driver maneuvers by considering holistic cues. With an array of sensors, driver's head, hand, and foot gestures are being captured in a synchronized manner together with lane, surrounding agents, and vehicle parameters. An emphasis is put on real-time algorithms. The cues are processed and fused using a latent-dynamic discriminative framework. As a case study, driver activity recognition and prediction in overtaking situations is performed using a naturalistic, on-road dataset. A consequence of this work would be in development of more effective driver analysis and assistance systems.**

I. INTRODUCTION

On-road driving behavior consists of intricate, multi-dimensional dynamics. It is the outcome of many variables that interact at a certain place and time. This motivates the study in this paper, which pursues a holistic approach using multiple cues in the scene. When fusing the different modalities in a temporal modeling framework, activity recognition and critical situation prediction can be made more effectively. Furthermore, through studying of temporally discriminative cues, we gain insight into the processes that characterize driver behavior.

One important implication of this work is in the development of driver assistance systems. For instance, human-observing cameras can perceive visual scanning and preparatory movements preceding a maneuver, giving a larger margin of time for prevention of a critical situation. Furthermore, a benefit of integrating driver gestures is in the ability to observe driver *intent* for performing a maneuver. Such knowledge can be incorporated into assitive technologies, which in turn may generate a more effective warning system under unintentional maneuvers (e.g. increased lane deviation), while assessing the need for a warning under intentional maneuvers. In this work, driver, vehicle, and environment cues are modeled to produce prediction of activities. In particular, overtaking event prediction will be studied to show the usefulness in the synergistic approach.

Many current safety systems sense the environment in order to produce a warning to the driver. We motivate the usage of additional cues-in particular, driver-based cues. In 2012 alone, 33,561 people died in motor vehicle traffic collisions in the United States [1]. A majority of such accidents involved an inappropriate maneuver or a distracted driver. Lateral control maneuvers such as overtaking and lane changing represent a significant portion of the total accidents each year. Between 2004-2008, 336,000 such crashes occurred in the US [2]. The major contributing factors are driver related (i.e. due to distraction or inappropriate decision making). Therefore, robust vision systems can be employed to detect driver motion patterns, such as head scanning or pre-control actions with foot or hand, and better mitigate critical situations. This work deals with such analysis in a

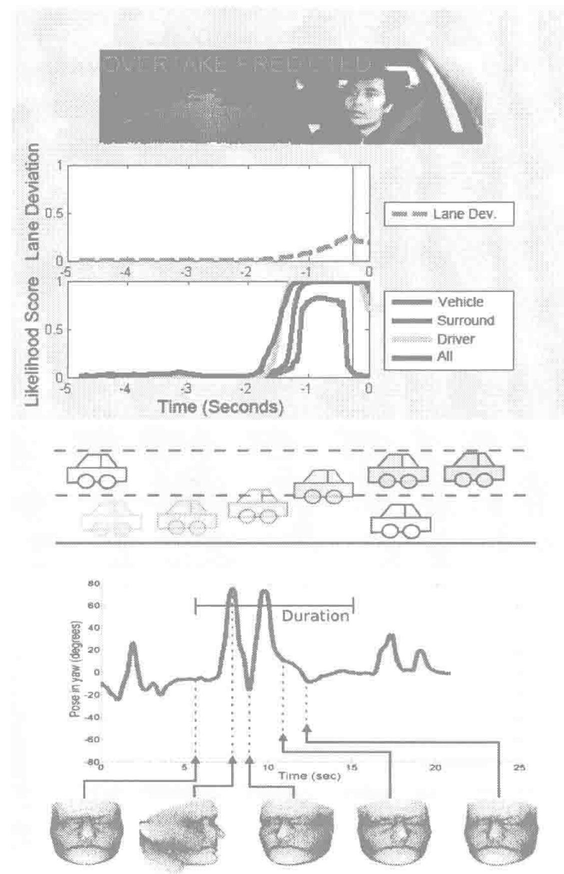

Fig. 1: Different cues are useful for analysis and prediction at different times of the maneuver. Top: An event from our dataset, with lane crossing marked in a vertical red line, and confidence scores from the model. Notice the spike occurring before the event, providing prediction. Bottom: Learning holistic features. For an overtake maneuver, visual scanning and head motion provides predictive value, especially when coupled with surround and vehicle dynamics cues.

E. Ohn-Bar, S. Martin, A. Tawari, and M. Trivedi are with the Laboratory for Intelligent and Safe Automobiles (LISA) at the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA {eohnbar, scmartin, atawari, mtrivedi}@ucsd.edu

978-1-4799-3639-7/14 $31.00 © 2014 IEEE

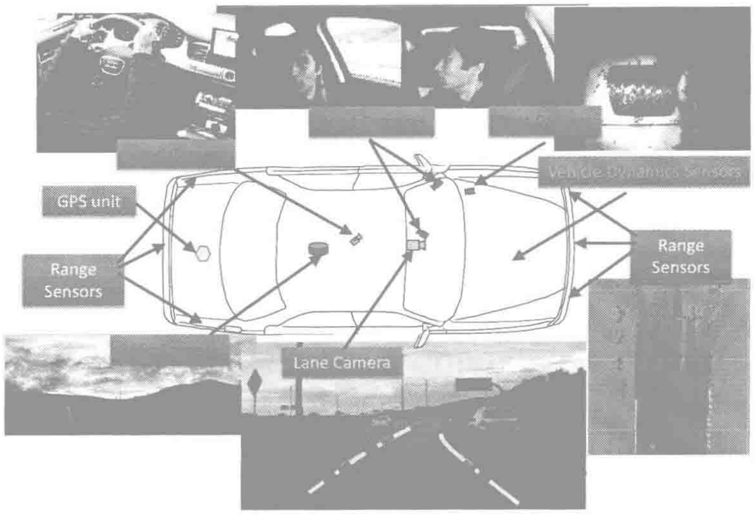

Fig. 2: Safety and driving behavior analysis using a multi-dimensional, holistic framework. The combined system can be used to predict events by a few seconds before they occur, thereby making roads safer and saving lives.

real-time action recognition framework.

First, in Section III we describe the testbed, and Section IV contains the vision algorithms used for signal extraction. The temporal formulation is details in Section V, and experimental analysis on the real-world driving dataset is in Section VII.

II. Related Research

Related research on maneuver recognition and prediction may differ based on cues used, the time-scale of the events of interest, or the type of maneuver studied.

In [3], vehicle dynamics measurements from the CAN-bus are coupled with laser data for preceding vehicles cues to predict driver behavior in a three second window. Yaw and steering-wheel angles were used in [4] for lane departure prediction. In [5] vehicle velocity, steering, and GPS were used to model lane-change behavior and predict collision trajectories. CAN parameters and a Gaussian Mixture Model were used in [6] to study car-following. Vehicle dynamics are used in [7] to identify driver intent at intersections and recognize actions. Vehicle trajectories were used in [8] to model driver intent at intersections using a Probabilistic Finite State Machine. Although methods commonly incorporate ego-vehicle dynamics, others may take on a purely vision approach. In [9], monocular video is used to predict driving behavior in urban environments. A front looking camera was used in [10] to predict future lane change behavior. Recognition of events was done using ego-vehicle dynamics and lane information in [11]. In our work, surrounding agents are modeled by using lidar cues, and a camera is used to extract lane parameters of the scene.

A close effort to our work can be found in [12], where radar, lane, head, and vehicle cues are integrated using a Relevance Vector Machine in order to recognize and predict lane change maneuvers in a two seconds window. A main difference is in the temporal modeling, as well as the cue representation. For instance, we produce a multi-level representation of the driver's state using head, hand, and foot motion cues. These cues provide additional context information for upcoming maneuvers. Furthermore, the temporal segmentation of actions is done automatically using the model.

Due to the holistic framework, it will be shown that prediction can be accomplished much earlier than in the aforementioned works. The event definition will be explored to highlight the earlier prediction. Generally, the aforementioned work predicts a maneuver after it has began using trajectory forecasting approaches. Nonetheless, we point out that the intent to perform the maneuver existed before the trajectory of the vehicle was altered and can be observed earlier. Within such an early time, a different set of cues may be useful for prediction, these are driver-based.

The closest work to ours of Doshi *et al.* [13] defines a lane change at the maximum lane deviation (i.e. when the vehicle crosses the lane marker). Nonetheless, the driver had the intent to change lanes much earlier, even before any lane deviation occurred. We therefore experiment with an alternate definition (used in [14]), which is at the beginning of the lateral movement. Both definitions will be studied in this work.

III. Experimental Setup: Testbed and Dataset

A 2011 Audi A8 was uniquely instrumented in order to capture the scene holistically: the dynamics of the ego-vehicle, surround agents and road geometry, and the driver's state. Fig. 2 shows a visualization of the sensor array, consisting of vision, radar, lidar, and vehicle CAN data. An on-board PC provides the computational resources. Sensor

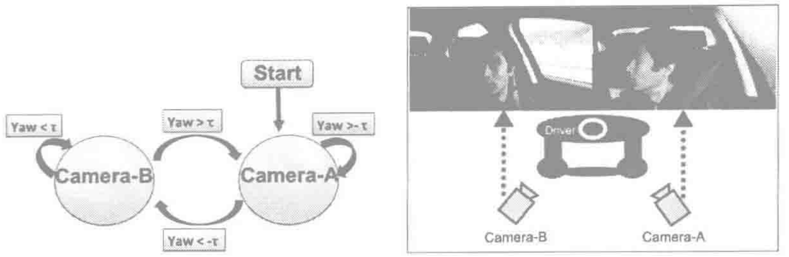

Fig. 3: A two camera system overcomes challenges in head pose estimation and allow for continuous tracking even under large head movements.

data from the radars and lidars are fused to generate a near-panoramic sensing of surrounding objects. These are tracked and associated using a module developed by Audi. UDP/TCP protocols are employed for synchronization of some of the sensors with the main PC. On our dataset, the sensors on average are synchronized up to 22ms.

The array of sensors includes vision-based, non-intrusive driver observing cameras. Two cameras for head pose tracking are used. These are aimed to capture head gestures, such as visual scanning. Two additional cameras provide inputs regarding gestures related to vehicle control. One camera is utilized for hand detection and tracking, and another camera for foot motion analysis.

For sensing the surround, a forward looking camera for lane tracking. Two lidar sensors, one forward and one facing backwards, and two radar sensors on either side of the vehicle. A Ladybug2 360° video camera (composed of an array of 6 individual rectilinear cameras) on top of the vehicle. Combined, these sensors allow for comprehensive and accurate on-line and off-line analysis and annotation.

The sensors are integrated into the vehicle body or placed in non-distracting regions to ensure minimal distraction while driving. Vehicle parameters are recorded into 13 measurements, such as steering angle, throttle and break, and vehicle's yaw rate.

In order to study the feasibility of the proposed framework, a dataset of 54 minutes of video containing 78,018 video frames was used (at 25 frames per second). All results reported employ 2-fold cross validation, with half of the data used for training and the rest for testing. Overall, we use 1000 normal events driving (each defined in a three second window leading to about 75,000 frames) as well as 13 overtaking instances (975 frames) which occurred throughout the video.

IV. Holistic Representation of Maneuvers

The features used to represent vehicle, surround, and driver, are detailed below. The output from each vision algorithm provides a time-series which is used in the temporal modeling for activity prediction. In this work, the panoramic 360° camera (composed of 6 individual streams) was used for annotation and offline analysis.

A. Head Pose Signal

Real-time, robust head pose is key to early prediction. The head provides a different set of cues (compared to hand and foot) because it is used by drivers for visual scanning and the retrieval of information from the surround. For instance, head motion may precede an overtaking maneuver in order to scan for an available space in the adjacent lane. On the other hand, controlling-gestures are associated with the foot and hand signals. These occur with the driver intention to operate a controller in the vehicle, such as in turning on a lane change indicator.

For capturing the wide motion of the head under such critical situations, we use a spatially distributed set of two cameras around the driver, as in [15].

First, head pose is estimated independently on each camera perspective from some of the least deformable facial landmarks (i.e. eye corners, nose tip), which are detected using supervised descent method [16], and their corresponding points on a 3D mean face model. The system runs at 50Hz. It is important to note that head pose estimation from each camera perspective is with respective camera coordinates. One-time calibration is performed to transform head pose estimation from respective camera coordinates to a common coordinate where a yaw rotation angle equal to, less than and greater than 0° represent the driver looking forward, rightward and leftward, respectively.

A simple camera selection module is used over the wide operational range in the yaw rotation angle, as shown in Fig. 3. In order to handle camera selection and hand-off, we use the yaw angle of the head.

B. Hand Location Signal

Hand gestures are incorporated in order to study preparatory motions before a maneuver is performed. Below, we specify the hand detection and tracking module. Hand detection is a difficult problem in computer vision, due to the hand tendency to occlude itself, deform, and rotate, producing a large variability in its appearance. In [17], motion and appearance cues were studied for hand activity classification in on-road data. As mentioned in [18], hand detection is particularly difficult in the car due to illumination variation. Therefore, we turned to training a hand detector on data from the same vehicle to maximize robustness to background artifacts. We use the fast to compute integral