Computer Vision and Image Understanding 134 (2015) 130–140

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems

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

Laboratory for Intelligent and Safe Automobiles (LISA), University of California, San Diego, CA 92093, USA

ARTICLE INFO

Article history:
Received 21 February 2014
Accepted 14 October 2014

Keywords:
Early activity recognition
Behavior-intent analysis
Multi-modal sensor fusion
Human–machine interactivity
Intelligent vehicles

ABSTRACT

We study techniques for monitoring and understanding real-world human activities, in particular of drivers, from distributed vision sensors. Real-time and early prediction of maneuvers is emphasized, specifically overtake and brake events. Study this particular domain is motivated by the fact that early knowledge of driver behavior, in concert with the dynamics of the vehicle and surrounding agents, can help to recognize dangerous situations. Furthermore, it can assist in developing effective warning and driver assistance systems. Multiple perspectives and modalities are captured and fused in order to achieve a comprehensive representation of the scene. Temporal activities are learned from a multi-camera head pose estimation module, hand and foot tracking, ego-vehicle parameters, lane and road geometry analysis, and surround vehicle trajectories. The system is evaluated on a challenging dataset of naturalistic driving in real-world settings.

1. Introduction

Distributed camera and sensor networks are needed for studying and monitoring agent activities in many domains of application [1]. Algorithms that reason over the multiple perspectives and fuse information have been developed with applications to outdoor or indoor surveillance [2]. In this work, multiple real-time systems are integrated in order to obtain temporal activity classification of video from a vehicular platform. The problem is related to other applications of video event recognition, as it requires a meaningful representation of the scene. Specifically, event definition and techniques for temporal representation, segmentation, and multi-modal fusion will be studied. These will be done with an emphasis on speed and reliability, which are necessary for the real-world, challenging application of preventing car accidents and making driving and roads safer. Furthermore, in the process of studying the usability and discriminative power of each of different cues, we gain insight into the underlying processes of driver behavior.

In 2012 alone, 33,561 people died in motor vehicle traffic crashes in the United States [3]. A majority of such accidents occurred due to an inappropriate maneuver or a distracted driver. In this work, we propose a real-time holistic framework for on-road analysis of driver behavior in naturalistic settings. Knowledge of the surround and vehicle dynamics, as well as the driver's state will allow the development of more efficient driver assistance systems. As a case study, we look into two specific maneuvers in order to evaluate the proposed framework. First, overtaking maneuvers will be studied. Lateral control maneuvers such as overtaking and lane changing represent a significant portion of the total accidents each year. Between 2004 and 2008, 336,000 such crashes occurred in the US [4]. Most of these occurred on a straight road at daylight, and most of the contribution factors were driver related (i.e. due to distraction or inappropriate decision making). Second, we look at braking events, which are associated with longitudinal control and their study also plays a key role in preventing accidents. Early recognition of dangerous events can aid in the development of effective warning systems. In this work we emphasize that the system must be extremely robust in order to: (1) engage only when it is needed by maintaining a low rate of false alarm rate, (2) function at a high true positive rate so that critical events, as rare as they may be, are not missed. In order to understand what the driver intends to do, a wide range of vision and vehicle sensors are employed to develop techniques that can satisfy real-world requirements.

The requirement for robustness and real-time performance motivates us to study feature *representation* as well as techniques for *recognition* of temporal events. The study will focus on three main components: the vehicle, the driver, and the surround. The implications of this study are numerous. In addition to early

* Corresponding author.
E-mail address: eohnbar@ucsd.edu (E. Ohn-Bar).

http://dx.doi.org/10.1016/j.cviu.2014.10.003
1077-3142/

warning systems, knowledge of the state of driver allows for customization of the system to the driver's needs, thereby mitigating further distraction caused by the system and easing user acceptance. On the contrary, a system which is not aware of the driver may cause annoyance. Additionally, under a dangerous situation (e.g. overtaking without turning on the blinker), a warning could be conveyed to other approaching vehicles. For instance the blinker may be turned on automatically.

Our goal is defined as follows: The prediction and early detection of overtaking and braking intent and maneuvers using driver, vehicle, and surround information.

In the vehicle domain, a few hundred milliseconds could signify an abnormal or dangerous event. To that end, we aim to model every piece of information suggesting an upcoming maneuver. In order detect head motion patterns associated with visual scanning [5–7] under settings of occlusion and large head motion, a two camera system for head tracking is employed. Subtle preparatory motion is studied using two additional cameras monitoring hand and foot motion. In addition to head, hand, and foot gesture analysis, sensors measuring vehicle parameters and surrounding vehicles are employed (Fig. 1). A gray-scale camera is placed in order to observe lane markings and road geometry, and a 360° color camera on top of the vehicle allows for panoramic analysis. Because visual challenges that are encountered in different surveillance domains, such as large illumination changes and occlusion, are common in our data, the action analysis modules studied in this work are generalizable to other domains of application as well.

We first perform a review of related literature in Section 2, while making a case for holistic understanding of multi-sensory fusion for the purpose of driver understanding and prediction. Event definition and testbed setup will be discussed in Sections 8 and 4, respectively. The different signals and feature extraction modules are detailed in Section 5. Two temporal modeling approaches for maneuver representation and fusion will be discussed in Section 6, and the experimental evaluation (Section 8) demonstrates analysis of different cues and modeling techniques in terms of their predictive power.

2. Related research studies

In our specific application, prediction involves recognition of distinct temporal cues not found in the large, 'normal' driving class. Related research may fall into three categories, which are roughly aligned with different temporal segments of the maneuver: trajectory estimation, inference, and intent prediction,with the first being the most common. In trajectory estimation, the driver is usually not observed, but IMU, GPS [8] and maps [9], vehicle dynamics [10], and surround sensors [11] play a role. These attempt to predict the trajectory of the vehicle given some observed evidence (i.e. the beginning of significant lateral motion) and the probability of crossing the lane marking [12,13]. A thorough recent review can be found in [14].

In intent inference approaches, the human is brought in as an additional cue, which may allow for earlier prediction. For instance, Doshi et al. [15] uses head pose, among other cues, in order to predict the probability that the vehicle will cross the lane marking in a two second window before the actual event. Several recent simulator studies have been performed using a variety of cues for intent inference. In [16], driver intent to perform overtaking was investigated using gaze information and an Artificial Neural Network (ANN). Vehicle dynamics, head, gaze, and upper body tracking cues were used in [17] with a rule-based approach for the analysis of driver intent to perform a variety of maneuvers. Even EEG cues may be used, as was done in [18] for emergency brake application prediction. Table 1 lists related research based on the maneuver studied, the learning approach, and the cues used for comparison with this work. Table 1 lists related studies done in naturalistic driving settings, as in our experiments. These present additional challenges to vision-based approaches.

Intent prediction corresponds to the earliest temporal prediction, and is rare in literature. Generally, existing studies do not look back in the prediction beyond 2–3 s before the event (e.g. the lane marker crossing for lane change maneuver). Intent prediction implies scene representation that may attempt to imitate human perception of the scene in order to produce a prediction for an intended maneuver. For instance, in [19] pre-attentive visual cues

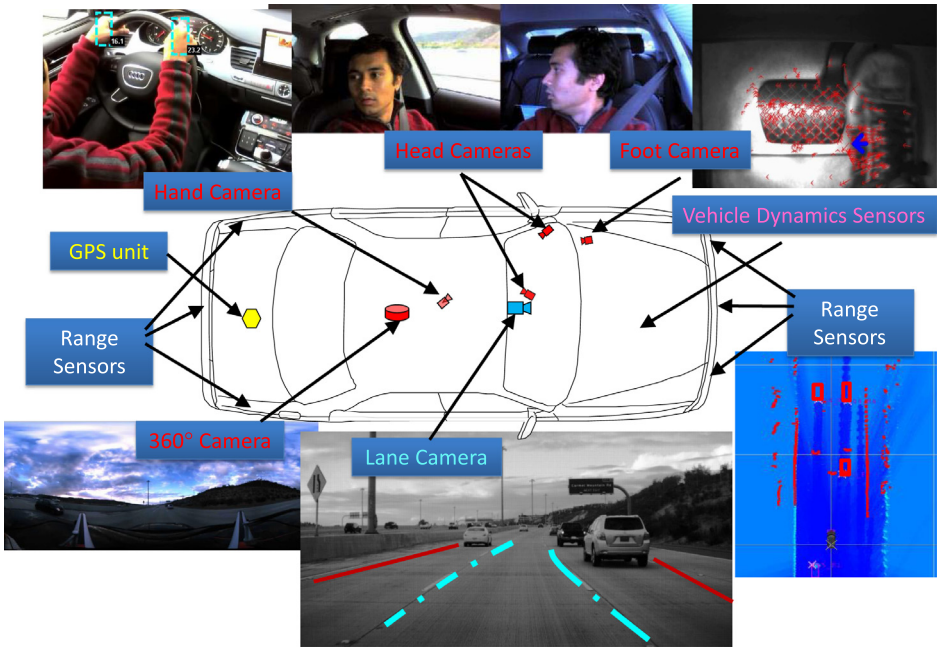


Fig. 1. Distributed, synchronized network of sensors used in this study. A holistic representation of the scene allows for prediction of driver maneuvers. Knowledge of events a few seconds before occurrence and the development of effective driver assistance systems could make roads safer and save lives.

Table 1
Overview of selected studies performed in real-world driving settings (i.e. as opposed to simulator settings) for maneuver analysis.

Study	Maneuvers	Inputs[a]	Method
McCall and Trivedi [20]	Brake	E, He, R, F	Relevance Vector Machine (RVM)
Doshi et al. [15]	Lane-change[b]	E, He, L, R	RVM
Tran et al. [21]	Brake	F	Hidden Markov Model (HMM)
Cheng et al. [22]	Turns	E, He, Ha	HMM
Pugeault and Bowden [19][c]	Brake, acceleration, clutch, steering	V	GIST + GentleBoost
Mori et al. [23]	Awareness during lane-change	R, Gaze	Correlation index
Liebner et al. [10]	Intersection turns and stop	GPS	Bayesian Network (BN)
Berndt and Dietmayer [24]	Lane-change and turns[b]	E, L, GPS, Map	HMM
This study[c]	Overtake, Brake	E, He, Ha, L, R, F, V	Latent-Dynamic Conditional Random Field (LDCRF) and Multiple Kernel Learning (MKL)

[a] Input types: E = Ego-vehicle parameters, He = Head, Ha = Hand, L = Lane, R = Radar/lidar objects, F = Foot, and V = Visual cues not included in previous types, such as break lights and pre-attentive cues.
[b] Defined lane-change at lane crossing.
[c] Explicitly models pre-intent cues.

from a front camera are learned for maneuver prediction. An example would be a brake light appearing in front of the ego-vehicle, causing the driver to brake.

In our objective to preform early prediction, we study a wide array of cues as shown in Table 1. In particular, we attempt to characterize maneuvers completely from beginning to end using both driver-based cues and surround-based cues. We point out that a main contribution comes from analysis of a large number of modalities combined, while other studies usually focused on a subset of the signals in this work (Table 1). Furthermore, the detection and tracking modules are all kept in real-time. Training and testing of models for intention prediction, inference, and trajectory estimation will be done. Furthermore, we study additional cues (hand, foot, visual pre-attentive cues) which were little studied in previous work. Studying driver, surround, and vehicle cues allows for gaining insight into how these are related throughout a maneuver (Fig. 2).

3. Event definition

Commonly, a lane change event or an overtake event (which includes a lane-change) are defined to begin at the lane marker crossing. On the contrary, in this work the beginning of an overtake event is defined earlier when the lateral motion started. We note that there are additional ways to define a maneuver such as an overtake or a lane-change (see [7]), and that our event start definition occurs significantly earlier than in many of the related research studies. For instance, techniques focusing on trajectory-based prediction define lane-change at the lane marker crossing.

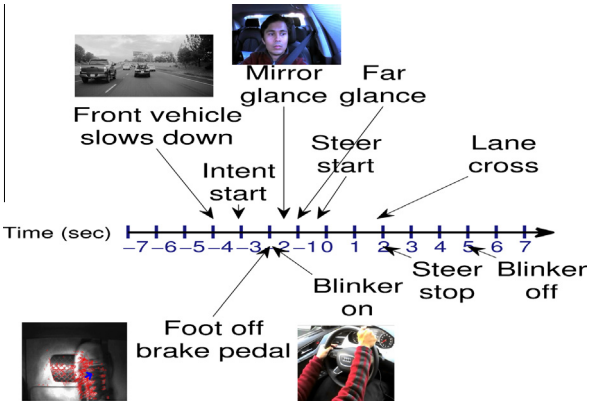


Fig. 2. Timeline of an example overtake maneuver. Our algorithm analyzes cues for intent prediction, intent inference, and trajectory estimation towards the end of the maneuver.

Nonetheless, as shown in (Fig. 2), the driver had the intent to change lanes much earlier, even before any lane deviation occurred. We wish to study how well can we observe such intent. By annotating events at the beginning of the lateral motion following the steering cue, the task of prediction becomes significantly more challenging. Under such a definition, lane deviation and vehicle dynamics are weak cues for prediction, while human-centered cues play a bigger role. Some examples are cues for visual scanning, as well as preparatory movements with foot and hands.

In addition to studying overtake maneuvers, which involve lateral control of the vehicle, we study a longitudinal control maneuver which is also essential in preventing accidents and monitoring for driver assistance. These are events where the driver chose to brake due to a situational need. While brakes are more easily defined (by pedal engagement), they allow us to evaluate the ability of the framework to generalize to other maneuvers. Any brake event (both harsh and weak) is kept in the data. This is done in order to emphasize analysis of key elements in the scene which cause drivers to brake.

4. Instrumented mobile testbed and dataset

A uniquely instrumented testbed vehicle is used in order to holistically capture the dynamics of the scene: the vehicle dynamics, a panoramic view of the surround, and the driver. Built on a 2011 Audi A8, the automotive testbed is outfitted with extensive auxiliary sensing for the research and development of advanced driver assistance technologies. Fig. 1 shows a visualization of the sensor array, consisting of vision, radar, lidar, and vehicle (CAN) data. The goal of the testbed buildup is to provide a near-panoramic sensing field of view for experimental data capture. Currently, the experimental testbed features robust computation in the form of a dedicated PC for development, which taps all available data from the on-board vehicle systems, excluding some of the camera systems which are synchronized using UDP/TCP protocols. Sensor data from the radars and lidars are fused into a single object list, with object tracking and re-identification handled by a sensor fusion module developed by Audi. On our dataset, the sensors are synchronized up to 22 ms (on average). The sensor list is as follows:

Looking into the vehicle:

- Two cameras for head pose tracking.
- One camera for hand detection and tracking.
- One camera for foot motion analysis.

Looking outside of the vehicle:

- Forward looking camera for lane tracking.

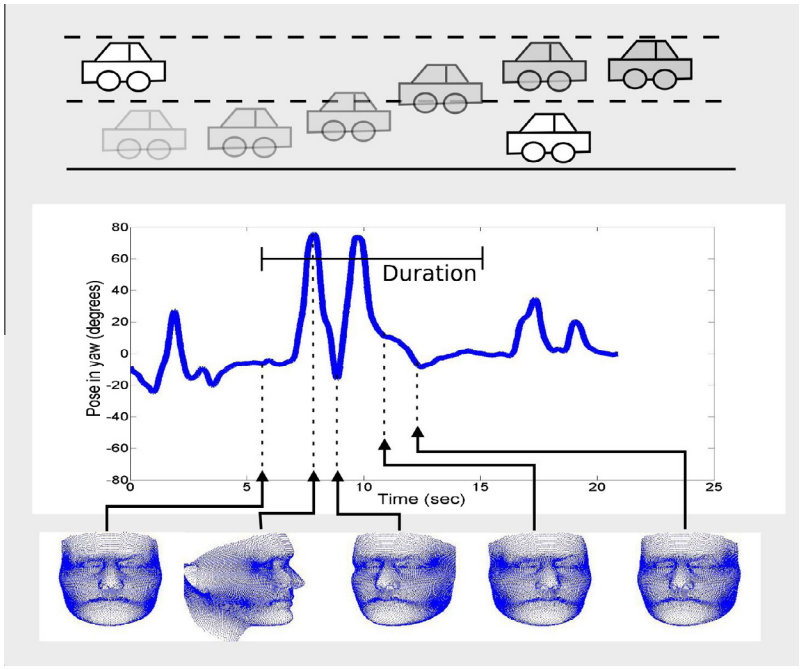


Fig. 3. An example overtake maneuver. Head cues are important for capturing visual scanning and observing intent. The output of the head pose tracker as the maneuver evolves are shown using a 3D model. See also Fig. 4

- Two lidar sensors, one forward and one facing backwards.
- 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.

The sensors are integrated into the vehicle body or placed in non-distracting regions to ensure minimal distraction while driving. Finally, information is captured from the CAN bus providing 13 measurements of the vehicle's dynamic state and controls, such as steering angle, throttle and brake, and vehicle's yaw rate.

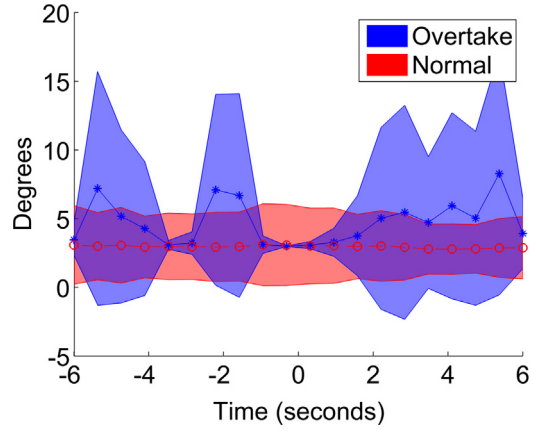
With this testbed, a dataset composed of three continuous videos with three different subjects for a total of about 110 min (over 165,000 video frames at 25 frames per second were used) was collected. Each driver was requested to drive as they would in naturalistic settings to a set of pre-determined set of destinations. Training and testing is done using a 3-fold cross validation over the different subjects, with two of the subjects used for training and the rest for testing. Overall, we randomly chose 3000 events of 'normal' driving with no brake or overtake events, 30 overtaking instances, and 87 brake events. Braking events may be harsh or soft, as any initial engagement of the pedal is used.

5. Maneuver representation

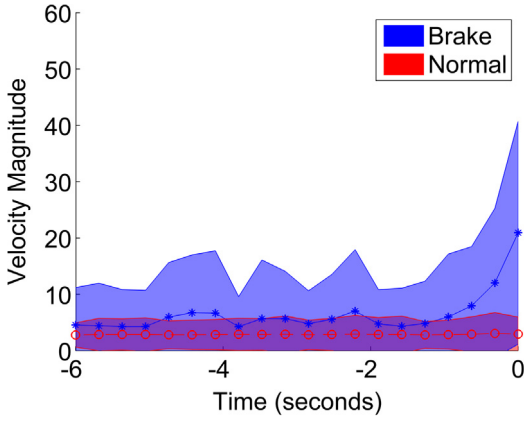
In this section we detail the vision modules used in order to extract useful signals for analysis of activities.

5.1. Signals

Head: Head dynamics are an important cue in prediction. The head differs from the other body parts since the head is used by drivers for information retrieval from the environment. For instance, head motion may precede an overtaking maneuver in order to scan for other vehicles (see Fig. 3). On the other hand, the foot and hand signals occur with the driver intention to operate a controller in the vehicle.



(a) Head yaw in degrees during an overtake event (t=0 at beginning of lateral motion).



(b) Foot velocity during a braking event.

Fig. 4. Mean and standard deviation of signals from the head pose and foot motion tracking modules during the two maneuvers studied in this paper.

Multiple cameras for human activity analysis [25] and face analysis [26] have been shown to reduce occlusion-related failures. In [27], a multi-perspective framework increased the operational range of monitoring head pose by mitigating failures under large head turns. In our setup, one camera is mounted on the front windshield near the A-pillar and another camera is mounted on the front windshield near the rear-view mirror to minimize intrusiveness.

First, head pose is estimated independently on each camera perspective using some of the least deformable facial landmarks (i.e. eye corners, nose tip), which are detected using supervised descent method [28], and their corresponding points on a 3D mean face model [29]. The system runs at 50 Hz. It is important to note that head pose estimation from each camera perspective is with respective to the 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.

Second, head pose is tracked over a wide operational range in the yaw rotation angle using both camera perspectives as shown in Fig. 5. In order to handle camera selection and hand-off, multiple techniques have been proposed in literature (a survey of different methods can be found at [1]). We had success with using the yaw as the camera hand-off cue. Assuming, without loss of generality, that at time $t = 0$ camera A is used to estimate head pose, then the switch to using camera B happens from when yaw rotation angle is greater than τ. Similarly the switch from B to A happens when yaw rotation angle is less than $-\tau$. If there is little to no spatial overlap in camera selection (i.e. $\tau = 0$), then noisy head pose measurements at the threshold will result in switching between the two camera perspectives needlessly. To avoid unnecessary switching between cameras, a sufficiently overlapping region is employed.

Hand: The hand signal will be used 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's tendency to occlude itself, deform, and rotate, producing a large variability in its appearance [30]. We use aggregate channel features [31] which are fast to extract. Specifically, for each patch extracted from a color image, gradient channels (six gradient orientation channels and normalized gradient magnitude) and color channels (CIE-LUV color channels were experimentally validated to work best compared to RGB and HSV) were extracted. 2438 instances of hands were annotated, and an AdaBoost classifier with decision trees as the weak classifiers is used for learning [32,33]. The hand detector runs at 30 fps on a CPU. We noticed many of the false detections occurring in the proximity of the actual hand (the arm, or multiple detections around the hand), hence we used a non-maximal suppression with a 0.2 threshold. Because of this, window size and padding had a significant effect on the results (Fig. 6). In order to differentiate the left from the right hand, we train a histogram of oriented gradients (HOG) with a support vector machine (SVM) detector. A Kalman filter is used for tracking.

Foot: One camera is used to observe the driver's foot behavior near the brake and throttle pedal, and an illuminator is also used due to lack of lighting in the pedal region. While embedded pedal sensors already exist to indicate when the driver is engaging any of the pedals, vision-based foot behavior analysis has additional benefits of providing foot movements before and after pedal press. Such analysis can be used to predict a pedal press before it is registered by the pedal sensors.

An optical flow (iterative pyramidal Lucas–Kanade [34], running at 30 Hz) motion cue is employed to determine the location of the foot and its velocity (see Fig. 7). Optical flow is sufficiently robust for analyzing foot behavior due to little illumination changes and the lack of other moving objects in the region. First, optical flow vectors are computed over sparse interest points, which are detected using Harris corner detection. Second, a majority vote over the computed flow vectors reveals an approximate location and magnitude of the global flow vector.

Optical flow based motion cues have been used in literature for analyzing head [35] and foot [21] gestures. Tran et al. [21] showed promising results where 74% of the pedal presses were correctly predicted 133 ms before the actual pedal press.

Lidar/radar: The maneuvers we study correlate with surrounding events. For instance, a driver may brake because of a forward vehicle slowing down or choose to overtake a vehicle in its proximity. Such cues are studied using an array of range sensors that track vehicles in term of their position and relative velocity. The sensor-fusion module, developed by Audi, tracks and re-identifies vehicles across the lidar and radar systems in a consistent global frame of reference. In this work we only consider trajectory information

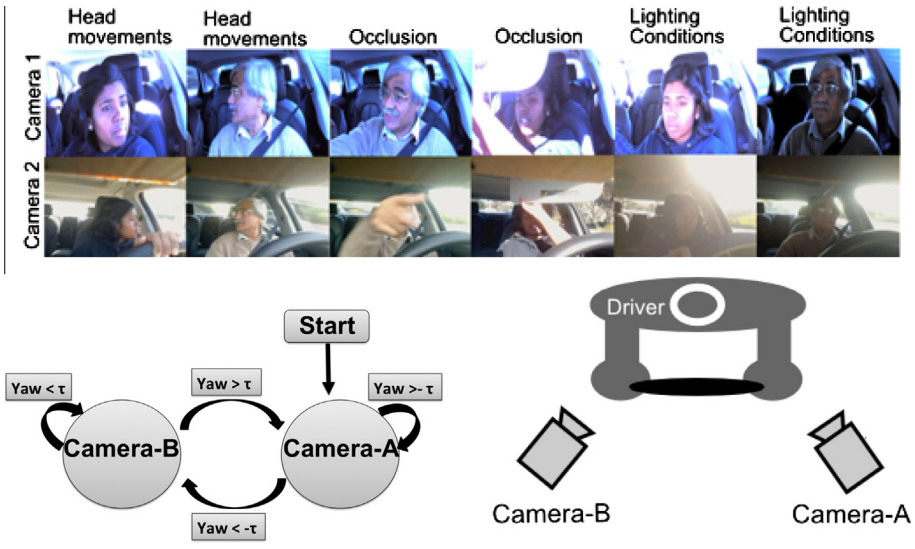


Fig. 5. A two camera system overcomes challenges in head pose estimation and allow for continuous tracking even under large head movements, varying illumination conditions, and occlusion.

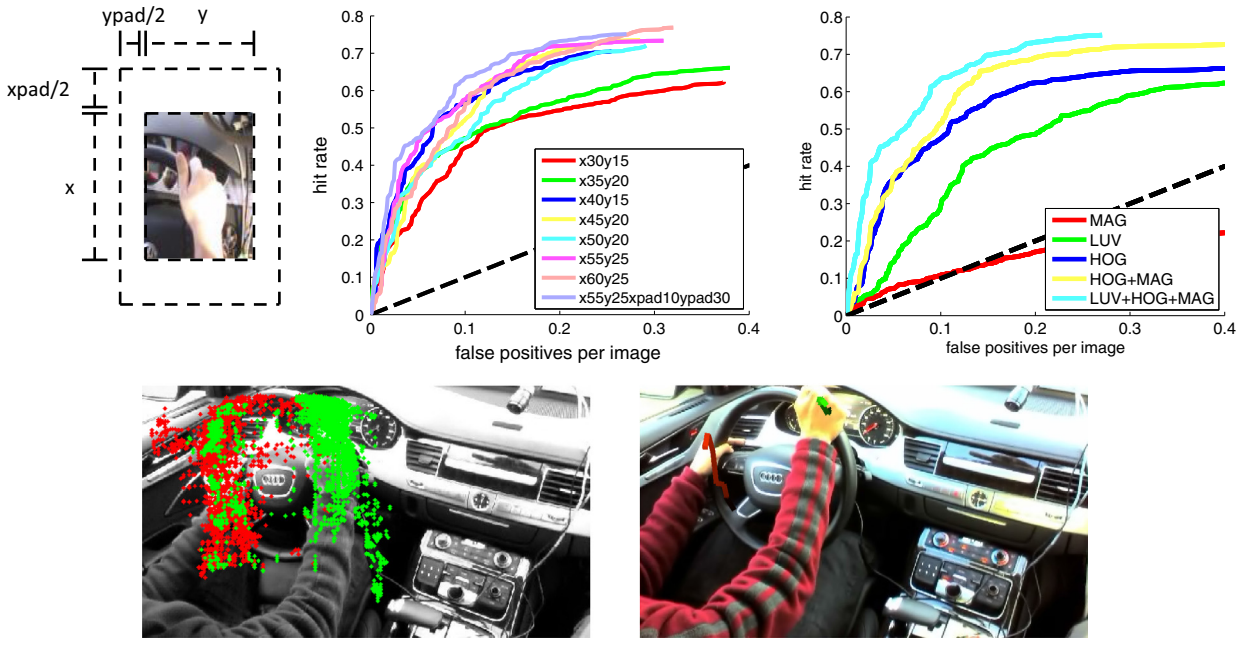


Fig. 6. Top: hand detection results with varying patch size and features; MAG – gradient magnitude, HOG – gradient orientation, and LUV color. Bottom: scatter plot of left (in red) and right (in green) hand detection for the entire drive. A hand trajectory of reaching towards the signal before an overtake is shown (brighter is later in time). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

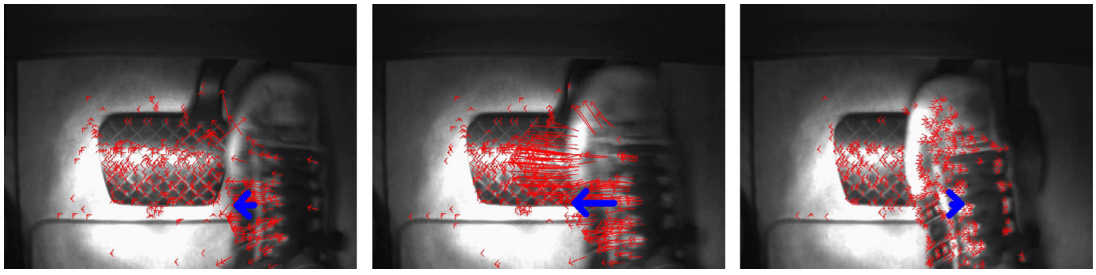


Fig. 7. Foot tracking using iterative pyramidal Lucas–Kanade optical flow. Majority vote produces location and velocity.

(longitudinal and lateral position and velocity) of the forward vehicle.

Lane: A front-observing gray-scale camera (see Fig. 1) is used for lane marker detection and tracking using a built-in system. The system can detect up to four lane boundaries. This includes the ego-vehicle's lanes and two adjacent lanes to those. The signals we consider are the vehicle's lateral deviation (position within the lane) and lane curvature.

Vehicle: The dynamic state of the vehicle is measured using a CAN bus, which supplies 13 parameters such as blinker state and vehicle's yaw rate. In understanding and predicting the maneuvers in this work, we only steering wheel angle information (important for analysis of overtake events), vehicle velocity, and brake and throttle pedal information.

Surround visual: The 360° panoramic camera outputs the composed view of six cameras. The view is used for annotation, offline analysis, as well as extracting color and visual information from the scene. The front vehicle, detected by the lidar sensor, is projected to the panorama image using an offline calibration. The projected vehicle box is padded, and a 50-bin histogram of the LUV channels is used as a descriptor for each frame. We also experimented with other scene descriptors, such as the GIST descriptor as done in [19]. GIST was shown to benefit cues that were not surround-observing (such as vehicle dynamics), yet the overall contribution after fusion of all of the sensors was not significant and so a detailed study of such features is left for future work.

5.2. Temporal features

We compare two temporal features for each of the signals outputted by any one of the sensors described above at each time, f_t. First, we simply use the signal in a time window of size L,

$$F_t = (f_{t-L+1}, \ldots, f_t) \tag{1}$$

The time window in our experiments is fixed at three seconds. These will be referred to as 'raw' features, as they simply involve a concatenation of the time series in the window.

A second set of features studied involves quantization of the signal into bins (states) in order to produce histograms (depicted in Fig. 8). The temporal feature is a normalized count of the states that occurred in the windowed signal. In this scheme, temporal information is preserved by a split of the signal into k equal sub-signals and histogram each of these sub-signals separately. We experimented with different choices for k, and found $k = 1, 2, 4$ to work well with no advantage in increasing the number of sub-segments further. This was used in all of the experiments. The number of bins was kept fixed at 20.

6. Temporal modeling

A model for the signals extracted by the modules in Section 5 must address several challenges. First, signal structure must be captured efficiently in order to produce a good modeling of maneuvers. Second, the role of different modalities should be studied with an appropriate fusion technique. Two types of modeling schemes are studied in this work, one using a Conditional Random Field (CRF) [36] and the other using Multiple Kernel Learning (MKL) [37]. The limitations and advantageous of these two schemes will be discussed, with the overarching goal of understanding the evolution and role of different signals in maneuver representation.

Given a sequence of observations from Eq. (1), $\mathbf{x} = \{F_t^{(1)}, \ldots, F_t^{(s)}\}$, where s is the total number of signals, the goal is to learn a mapping to a label space, $\mathcal{Y}$, of different maneuver labels. This can be done using a conditional random field.

Conditional random field: Temporal dynamics are often modeled using a graphical model which reasons over the temporal structure of the signal. This can be done by learning a generative model, such as a Markov Model (MM) [22], or a discriminative model such as a Conditional Random Field (CRF) [36]. Generally, CRF has been shown to significantly outperform its generative counterpart, the MM. Furthermore, CRF can be modified to better model latent temporal structures, which is essential for our purposes.

The Hidden CRF (HCRF) [38] introduces hidden states that are coupled with the observations for better modeling of parts in the temporal structure of a signal with a particular label. A similar mechanism is employed by the Latent-Dynamic CRF (LDCRF) [36], with the advantage of also providing a segmentation solution for a continuous data stream. Defining a latent conditional model and assuming that each class label has a disjoint set of associated hidden states $\mathbf{h}$ gives

$$P(\mathbf{y}|\mathbf{x};\Lambda) = \sum_{\mathbf{h}} P(\mathbf{y}|\mathbf{h},\mathbf{x},\Lambda)P(\mathbf{h}|\mathbf{x},\Lambda) = \sum_{\mathbf{h}:\forall h_i \in H_{y_i}} P(\mathbf{h}|\mathbf{x};\Lambda) \tag{2}$$

where Λ is the set of model parameters and $\mathbf{y}$ is a label or a sequence of labels. In a CRF with a simple chain assumption, this joint distribution over $\mathbf{h}$ has an exponential form,

$$P(\mathbf{h}|\mathbf{x};\Lambda) = \frac{\exp(\sum_k \Lambda_k \cdot \mathbf{T}_k(\mathbf{h},\mathbf{x}))}{\sum_{\mathbf{h}} \exp(\sum_k \Lambda_k \cdot \mathbf{T}_k(\mathbf{h},\mathbf{x}))} \tag{3}$$

We follow [36], where the function $\mathbf{T}_k$ is defined as a sum of state (vertex) or binary transition (edge) feature functions,

$$\mathbf{T}_k(\mathbf{h},\mathbf{x}) = \sum_{i=1}^{m} l_k(h_{i-1}, h_i, \mathbf{x}, i) \tag{4}$$

The model parameters are learned with gradient ascent over the training data using the objective function,

$$L(\Lambda) = \sum_{i}^{n} \log P(\mathbf{y}_i|\mathbf{x}_i,\Lambda) - \frac{1}{2\sigma^2}||\Lambda||^2 \tag{5}$$

where $P(\Lambda) \sim \exp(\frac{1}{2\sigma^2}||\Lambda||^2)$. In inference, the most probable sequence of labels is the one that maximizes the conditional model (Eq. (2)). Marginalization over the hidden states is computed using belief propagation.

With LDCRF, early-fusion is used for fusion of the temporal signal features. When considering the histogram features studied in this work, each bin in the histogram is associated with an observation vector of size k (where k is illustrated in Fig. 8). In this case, temporal structure is measured by the evolution of each bin over time. Possibly due to the increase in dimensionality and the already explicit modeling of temporal structure in the LDCRF model, using raw features was shown to work as good or better than the sub-segment histogram features.

Multiple kernel learning: A second approach for constructing a maneuver model is motivated by the need for fusion of the large number of incoming signals from a variety of modalities. Given a set of training instances and signal channel c_l (i.e. brake pedal output), a kernel function is calculated for the signal, $\kappa_{c_l}(\mathbf{x}_i,\mathbf{x}_j) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (d is the feature dimension and $\mathbf{x}_i,\mathbf{x}_j$ are two data points). This produces a set of s kernel matrices for the n data points in the training set, $\{\mathbf{K}^{c_l} \in \mathbb{R}^n \times \mathbb{R}^n, l = 1,\ldots,s\}$, so that $K_{ij}^{c_l} = \kappa_{c_l}(\mathbf{x}_i,\mathbf{x}_j)$. s stands for the total number of outputs provided by the modules in Section 5. In our implementation, Radial Basis Function (RBF) kernels are derived for each of the signals using $\kappa(\mathbf{x}_i,\mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||/\gamma)$. The cost and spread parameters are found for each signal separately using grid search.

The kernels are combined by learning a probability distribution $\mathbf{p} = (p^1,\ldots,p^s)$, with $p^l \in \mathbb{R}_+$ and $\mathbf{p}^T\mathbf{1} = 1$, such that the combination of kernel matrices,

$$\mathbf{K}(\mathbf{p}) = \sum_{l=1}^{s} p^l \mathbf{K}^{c_l} \tag{6}$$

is optimal. In this work, the weights are learned using stochastic approximation [37]. LIBSVM [39] is used as the final classifier. The histogram features were shown to work well with MKL, performing better than simply using the raw temporal signal features [40].

7. Experimental setup

Several experiments are conducted in order to test the proposed framework for recognition of intent and prediction of maneuvers. As mentioned in Section 3, we experiment with two definitions for the beginning of an overtake event. An overtake event may be marked when the vehicle crossed the lane marking or when the lateral movement began. These are referred to as overtake-late and overtake-early, respectively. Normal driving is defined as

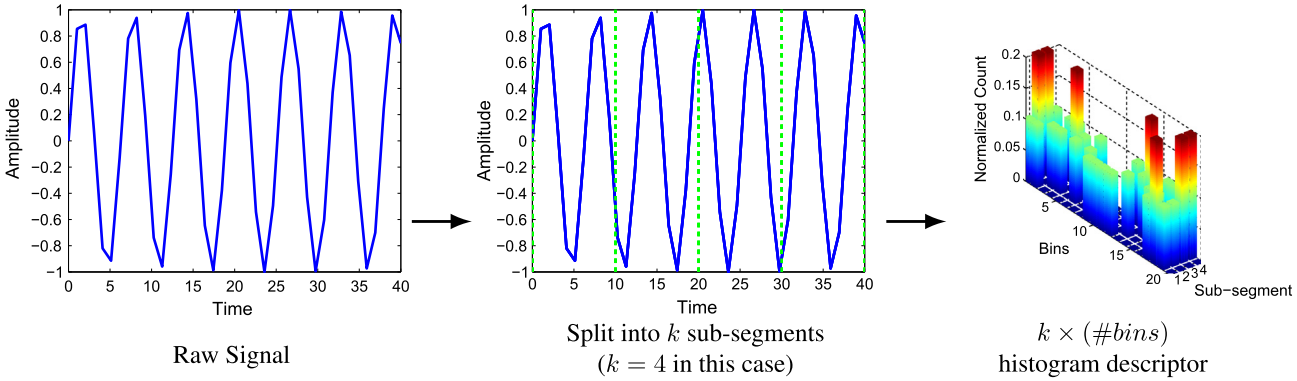


Fig. 8. Two features used in this work: raw trajectory features outputted by the detection and tracking, and histograms of sub-segments.

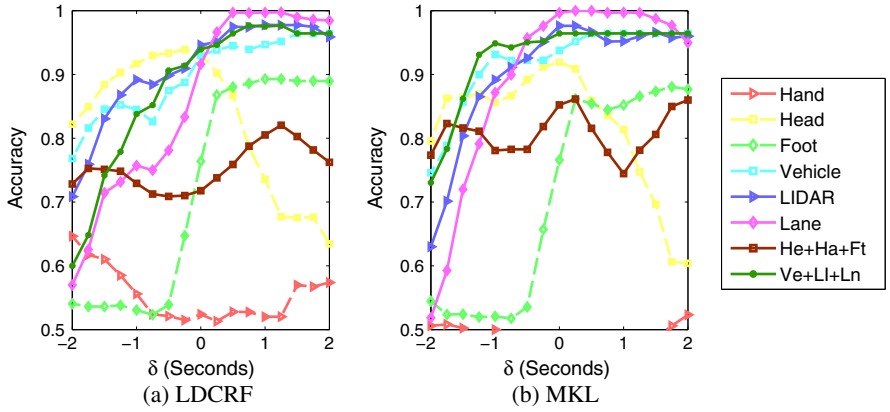


Fig. 9. Classification and prediction of overtake-late/brake (Experiment 1a) maneuvers using raw trajectory features. He + Ha + Ft stands for the driver observing cues head, hand, and foot. Ve + Li + La is vehicle (CAN), lidar, and lane. MKL is shown to handle integration of multiple cues better.

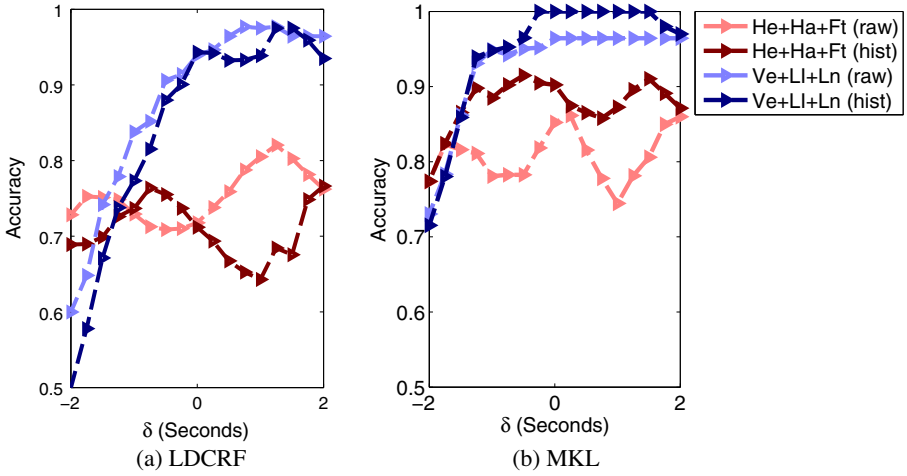


Fig. 10. Comparison of the two temporal features (see Section 5.2) studied in this work, raw temporal features and sub-segments histogram features, using overtake-late/brake (Experiment 1a) maneuvers. MKL benefits from the histogram features, especially in fusion of multiple types of modalities.

events when the brake pedal was not engaged and no significant lane deviation occurred, but the driver was simply keeping within the lanes. A brake event is any event in which the brake pedal became engaged. Furthermore, we do not require a minimum speed for the events, so normal, brake, and overtake events may occur at any speed. Brake events may be in any magnitude of pedal press.

Initially, the proposed framework is evaluated by studying the question of whether a driver is about to overtake of brake due to a leading vehicle, as both are possible maneuvers. These experiments provide analysis on the temporal features and modeling. Once these initial experiments are complete, this allows us to move further to more complicated scenarios. Below, we detail the reference system to each experiment that will be performed in the experimental evaluation (Section 8).

- *Experiment 1a:* Overtake-late events vs. brake events (overtake-late/brake).
- *Experiment 1b:* Overtake-early events vs. brake events (overtake-early/brake).

Next, we are concerned with how each of the above events is characterized compared to normal driving.

- *Experiment 2a:* Overtake-late events vs. normal driving events (overtake-late/normal).
- *Experiment 2b:* Overtake-early events vs. normal driving events (overtake-early/normal).

Finally, we study the framework under a different maneuver,

- *Experiment 3:* Brake events vs. normal driving (brake/normal).

8. Experimental evaluation

Temporal modeling: The first set of evaluations is concerned with comparison among the choices for the temporal features and temporal modeling. Each cue is first modeled independently in order to study its predictive power. The results for LDCRF and MKL under experiment 1a, overtake-late/brake are shown in Fig. 9 for raw trajectory features. LDCRF demonstrates better predictive power using each modality independently when compared to MKL. For instance, lane information provides better prediction at $\delta = -2$ (2 s before the event start definition) with the LDCRF model. Similar conclusion holds for the head pose signal as well.

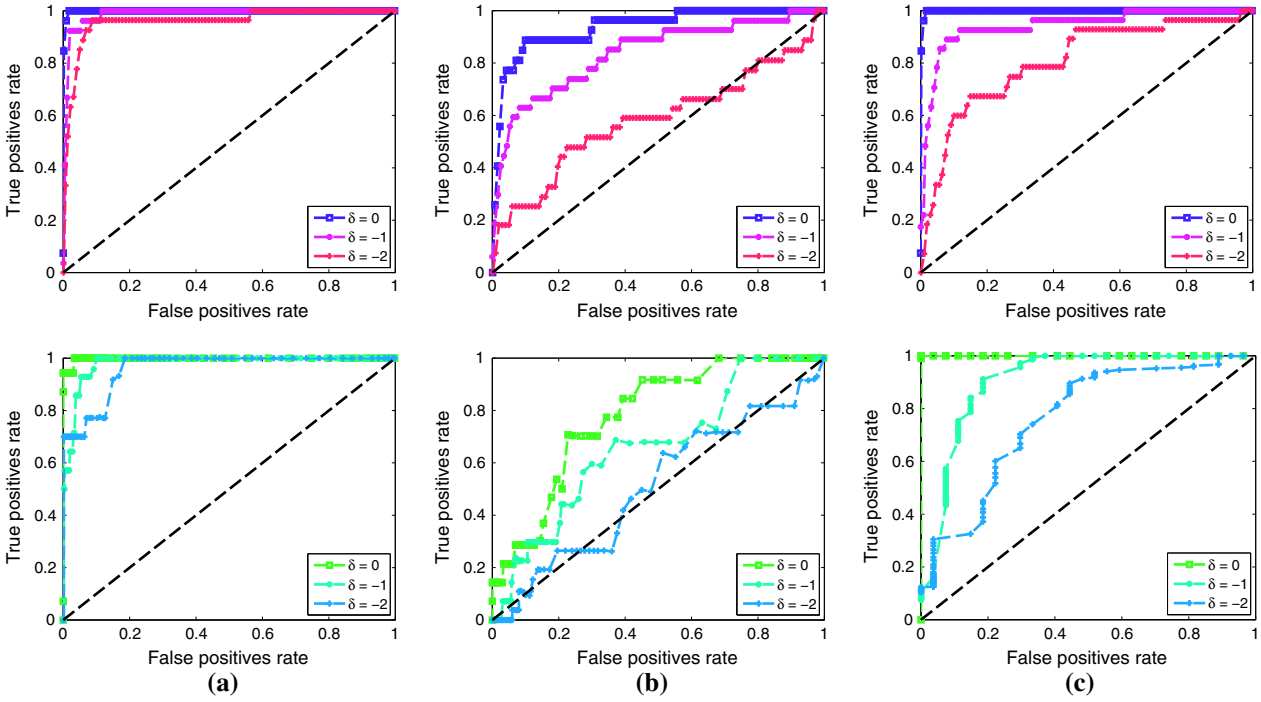


Fig. 11. Measuring prediction by varying the time in seconds before an event, δ. Top: MKL results. Bottom: LDCRF results. (a) Experiment 2a: overtake-late vs. normal (b) Experiment 2b: overtake-early vs. normal (c) Experiment 3: brake vs. normal. Note how prediction of overtake-early events, which occur seconds before the beginning of an overtake-late events, is more difficult.

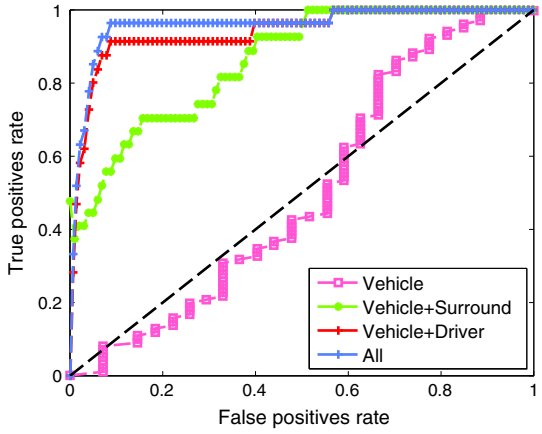


Fig. 12. For a fixed prediction time of $\delta = -2$ s, we show the effects of appending cues to the vehicle dynamics under overtake-late/normal (Experiment 2a). The surround cues utilize lidar, lane, and visual data. Driver cues include the hand, head, and foot signals.

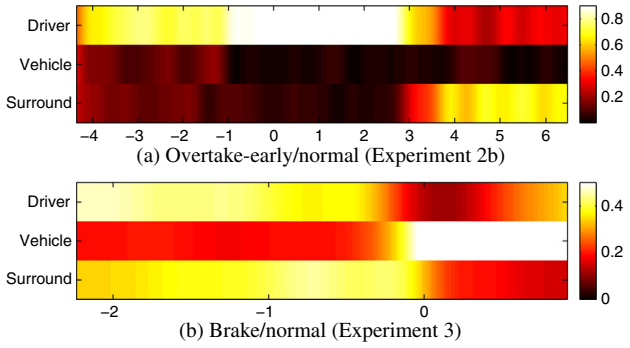


Fig. 13. Kernel weight associated with each cue learned from the dataset with MKL (each column sums up to one). Each manuever was learned against a set of normal events without the maneuver. Characterizing a maneuver requires cues from the human (hand, head, and foot), vehicle (CAN), and the environment (lidar, lane, visual-color changes). Time 0 for overtake is at the beginning of the lateral motion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As LDCRF explicitly reasons over temporal structure in the signal, these results are somewhat expected.

Temporal features and fusion: Fig. 9 also shows the results of fusion of multiple modalities with one model learned over the multiple types of signals. For clarity, we only show fusion of driver-based cues (head, hand, and foot) and surround cues (vehicle parameters, lidar, and lane). MKL is shown to perform better, as it is designed for fusion of multiple sources of signals. On the other hand, with the increase in dimensionality, the LDCRF model is shown to be limited. This is further studied in Fig. 10, where the MKL scheme demonstrates further gains due to the temporal structure encoded by the histogram descriptor. This is not the case for LDCRF, as it already explicitly reasons over temporal structure in the data. Therefore, for the rest of the section, LDCRF is joined with raw temporal features and the MKL with the temporal histogram features. Next, the more challenging experiments of early prediction are performed. As specific events are studied against a large 'normal' events dataset which includes naturalistic variation in each cue, the prediction task becomes more challenging. Furthermore, prediction much earlier in the maneuver of overtake-early events is also challenging.

The results are summarized in Fig. 11 for experiments 2 and 3, where the entire set of signals described in Section 5 is used. For each experiment, the predictive power of the learned model is measured by making a prediction of a maneuver earlier in time, at increments of one second. At $\delta = -2$, a prediction is made two seconds before the actual event definition. Fig. 11(b) demonstrates the challenging task of prediction of overtake-early events, which mostly involve recognition of scanning and preparatory movement together with the surround cues. In this scenario of intent inference, lane deviation or steering angle info (which are strong cues

for prediction in overtake-late events) are less informative. On the other hand, prediction of two seconds before an overtake-late maneuver is well defined in the feature space. Generally, the MKL is shown better results due to better fusion of the multiple signal sources, yet the prediction trends are consistent with the two temporal modeling schemes.

Insights into the maneuvers: Next, we consider the trade-off and value in sensor addition to an existing vehicle system. Suppose that vehicle dynamics are provided, we quantify the benefit of adding a surround sensor capturing system for the prediction compared to a driver sensing system. The results are depicted in Fig. 12. Although both systems provide an advantage, most gains for early prediction come for prediction by observing driver related cues.

Fig. 13 shows the temporal evolution of cue importance using the weight output $\mathbf{p}$ from the MKL framework. Effective kernels will correspond to a heavier weight, and kernels with little discriminative value will be associated a smaller weight. Fig. 13 demonstrates how the entire maneuver can now be characterized in terms of the dynamics and evolution of different cue over the maneuver. For overtake events, driver-related cues of head, hand, and foot are strongest around the time that the lateral motion begins ($t = 0$) in Fig. 13(a). Surround cues include lane, lidar, and visual surround cues. After the steering began, the lane deviation cue becomes a strong indicator for the activity. Similarly, the temporal evolution of the cues is shown for brake/normal event classification in Fig. 13(b). We see that driver cues (i.e. foot), and surround cues (i.e. visual cues, lidar) are best for early prediction, and a sharp increase in the kernel weight associated with vehicle dynamics occurs around the time of the pedal press.

9. Concluding remarks

In this work, a surveillance application of driver assistance was studied. Automotive driver assistance systems must perform under time-critical constraints, where even tens of milliseconds are essential. A holistic and comprehensive understanding of the driver's intentions can help in gaining crucial time and save lives. Prediction of human activities was studied using information fusion from an array of sensors in order to fully capture the development of complex temporal interdependencies in the scene. Evaluation was performed on a rich and diverse naturalistic driving dataset showing promising results for prediction of both overtaking and braking maneuvers. The framework allowed the study of the different types of signals over time in terms of predictive importance. In the future, additional maneuver types, such as those performed when approaching to and at intersections will be studied.

Acknowledgments

The authors would like to thank the reviewers and editors for their helpful comments. The authors gratefully acknowledge sponsorship of the UC Discovery Program and associated industry partners including Audi, Volkswagen Electronics Research Laboratory, and Toyota Motors. Support of colleagues from the UCSD Laboratory for Intelligent and Safe Automobiles is also appreciated.

References

[1] B. Bhanu, C.V. Ravishankar, A.K. Roy-Chowdhury, H. Aghajan, D. Terzopoulos (Eds.), Distributed Video Sensor Networks, Springer, 2011.
[2] S. Calderara, A. Prati, R. Cucchiara, Hecol: Homography and epipolar-based consistent labeling for outdoor park surveillance, Comput. Vision Image Understand. 111 (2008) 21–42.
[3] 2012 Motor Vehicle Crashes: Overview, Tech. Rep. DOT HS 811 856, National Highway Traffic Safety Administration, Washington, DC, 2013.
[4] W.G. Najm, R. Ranganathan, G. Srinivasan, J.D. Smith, S. Toma, E. Swanson, A. Burgett, Description of Light-vehicle Pre-crash Scenarios for Safety Applications based on Vehicle-to-vehicle Communications, Tech. Rep. DOT HS 811 731, National Highway Traffic Safety Administration, Washington, DC, 2013.
[5] P.M. Valero-Moraa, A. Tontscha, R. Welshb, A. Morrisb, S. Reedb, K. Touliouc, D. Margaritisc, Is naturalistic driving research possible with highly instrumented cars? Lessons learnt in three research centres, Accid. Anal. Prev. 58 (2013) 187–194.
[6] T. Taylora, A. Pradhanb, G. Divekara, M. Romosera, J. Muttarta, R. Gomeza, A. Pollatsekc, D. Fisherd, The view from the road: the contribution of on-road glance-monitoring technologies to understanding driver behavior, Accid. Anal. Prev. 58 (2013) 175–186.
[7] A Comprehensive Examination of Naturalistic Lane-changes, Tech. Rep. DOT HS 809 702, National Highway Traffic Safety Administration, Washington, DC, 2004.
[8] R. Simmons, B. Browning, Y. Zhang, V. Sadekar, Learning to predict driver route and destination intent, in: IEEE Conf. Intelligent Transportation Systems, 2006.
[9] S. Lefèvre, C. Laugier, J. Iba nez-Guzmán, Exploiting map information for driver intention estimation at road intersections, in: IEEE Intelligent Vehicles Symposium, 2011.
[10] M. Liebner, M. Baumann, F. Klanner, C. Stiller, Driver intent inference at urban intersections using the intelligent driver model, in: IEEE Intelligent Vehicles Symposium, 2012.
[11] M. Ortiz, F. Kummert, J. Schmudderich, Prediction of driver behavior on a limited sensory setting, in: IEEE Conf. Intelligent Transportation Systems, 2012.
[12] V. Gadepally, A. Krishnamurthy, Ü. Özgüner, A Framework for Estimating Driver Decisions Near Intersections, IEEE Trans. Intell. Transp. Syst. 15 (2) (2014).
[13] S. Lefèvre and J. Ibañez-Guzmán and C. Laugier, IEEE Symposium on Comp. Intell. Veh. Transp. Syst. (2011)
[14] A. Doshi, M.M. Trivedi, Tactical driver behavior prediction and intent inference: a review, in: IEEE Conf. Intelligent Transportation Systems, 2011.
[15] A. Doshi, B.T. Morris, M.M. Trivedi, On-road prediction of driver's intent with multimodal sensory cues, IEEE Pervasive Comput. 10 (2011) 22–34.
[16] F. Lethaus, M.R. Baumann, F. Kster, K. Lemmer, A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data, Neurocomputing 121 (0) (2013) 108–130.
[17] M.-I. Toma, D. Datcu, Determining car driver interaction intent through analysis of behavior patterns, in: Technological Innovation for Value Creation, Springer, 2012, pp. 113–120.
[18] S. Haufe, M.S. Treder, M.F. Gugler, M. Sagebaum, G. Curio, B. Blankertz, EEG potentials predict upcoming emergency brakings during simulated driving, J. Neural Eng. 8 (2011) 056001.
[19] N. Pugeault, R. Bowden, Learning pre-attentive driving behaviour from holistic visual features, in: European Conf. Computer Vision, 2010.
[20] J. McCall, M.M. Trivedi, Driver behavior and situation aware brake assistance for intelligent vehicles, Proc. IEEE 95 (2007) 374–387.
[21] C. Tran, A. Doshi, M.M. Trivedi, Modeling and prediction of driver behavior by foot gesture analysis, Comput. Vision Image Understand. 116 (2012) 435–445.
[22] S.Y. Cheng, S. Park, M.M. Trivedi, Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis, Comput. Vision Image Understand. 106 (2007) 245–257.
[23] M. Mori, C. Miyajima, P. Angkititrakul, T. Hirayama, Y. Li, N. Kitaoka, K. Takeda, Measuring driver awareness based on correlation between gaze behavior and risks of surrounding vehicles, in: IEEE Conf. Intelligent Transportation Systems, 2012.
[24] H. Berndt, K. Dietmayer, Driver intention inference with vehicle onboard sensors, in: IEEE Conf. Vehicular Electronics and Safety, 2009.
[25] R. Cucchiara, A. Prati, R. Vezzani, A multi-camera vision system for fall detection and alarm generation, Exp. Syst. 24 (2007) 334–345.
[26] L. An, M. Kafai, B. Bhanu, Dynamic bayesian network for unconstrained face recognition in surveillance camera networks, IEEE Trans. Emerg. Sel. Top. Circ. Syst. 3 (2) (2013) 155–164.
[27] A. Tawari, S. Martin, M.M. Trivedi, Continuous head movement estimator (CoHMET) for driver assistance: issues, algorithms and on-road evaluations, IEEE Trans. Intell. Transport. Syst. 15 (2014) 818–830.
[28] X. Xiong, F.D.I. Torre, Supervised descent method and its application to face alignment, in: IEEE Conf. Computer Vision and Pattern Recognition, 2013.
[29] S. Martin, A. Tawari, E. Murphy-Chutorian, S.Y. Cheng, M. Trivedi, On the design and evaluation of robust head pose for visual user interfaces: algorithms, databases, and comparisons, in: ACM Conf. Automotive User Interfaces and Interactive Vehicular Applications, 2012.
[30] E. Ohn-Bar, S. Martin, M.M. Trivedi, Driver hand activity analysis in naturalistic driving studies: issues, J Electron. Imaing 22 (4) (2013) 1–10.
[31] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Machine Intell. 36 (8) (2014) 1532–1545.
[32] C. Zhang, P.A. Viola, Multiple-instance pruning for learning efficient cascade detectors, in: Advances in Neural Information Processing Systems, 2007.
[33] E. Ohn-Bar and M.M. Trivedi, Beyond just keeping hands on the wheel: Towards visual interpretation of driver hand motion patterns, in: IEEE Conf. Intelligent Transportation Systems, 2014.
[34] J. Bouguet, Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. Intel Corporation (2001).
[35] S. Martin, C. Tran, M. Trivedi, Optical flow based head movement and gesture analyzer (OHMeGA), in: IEEE Intl. Conf. Pattern Recognition, 2012.
[36] L.P. Morency, A. Quattoni, T. Darrell, Latent-dynamic discriminative models for continuous gesture recognition, in: IEEE Conf. Computer Vision and Pattern Recognition, 2007.

[37] S. Bucak, R. Jin, A.K. Jain, Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition, in: Advances in Neural Information Processing Systems, 2010.
[38] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell, Hidden conditional random fields, IEEE Trans. Pattern Anal. Machine Intell. 29 (2007) 1848–1853.
[39] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.
[40] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, Vision on wheels: Looking at driver, vehicle, and surround for on-road Maneuver analysis, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014.