IEEE Conference on Intelligent Transportation Systems 2014 (to appear)

Beyond Just Keeping Hands on the Wheel: Towards Visual Interpretation of Driver Hand Motion Patterns

Eshed Ohn-Bar and Mohan M. Trivedi†

***Abstract*— Observing hand activity in the car provides a rich set of patterns relating to vehicle maneuvering, secondary tasks, driver distraction, and driver intent inference. This work strives to develop a vision-based framework for analyzing such patterns in real-time. First, hands are detected and tracked from a monocular camera. This provides position information of the left and right hands with no intrusion over long, naturalistic drives. Second, the motion trajectories are studied in settings of activity recognition, prediction, and higher-level semantic categorization.**

I. Introduction

This study is concerned with construction of robust, vision-based tools for studying hand motion patterns under naturalistic, real-world settings. Since the study of human hands is an active field in the computer vision, machine learning, and human-machine interaction communities, the methods developed in this work are relevant to a wide array of applications. Inferring hand activity is especially important in the operated vehicle, as hands are a common medium for expressing and conveying information. For instance, it may provide vital information about the state of attentiveness of the driver. In order to clearly motivate the study, we list potential applications below.

Motivating applications: First, hand tracking allows the study of *preparatory movements* for maneuvers [1], [2]. Such information may be useful when providing alerts and support to the driver [3]. For instance, while performing a sharp turn a driver may shift the hand position while the turn is ongoing in order to further turn the wheel, an action which may lead to an accident. Another example is in preparing for an overtaking maneuver, where a driver may shift the hand position together with a sequence of head and body pose dynamics to prepare for the overtake [2]. A second potential application is in *monitoring distraction levels*, as hand-vehicle and hand-object interactions (such as text messaging, handling navigation, etc.) can potentially increase visual, manual, and cognitive load [4]. This important application is presseing as drivers today are increasingly engaged in secondary tasks behind the wheel (23.5% of the time according to [4]). A third possible application lies in providing a framework for hand gesture recognition for *interactivity*, as in [5]. Finally, *long term analysis* of hand motion can provide useful insight into crash and near-crash events. For instance, in studying gestures performed by the driver for re-gaining control following an unexpected event. The framework proposed in this paper can be immediately applied to other applications of hand gesture recognition [6], such as tutoring applications as in [7].

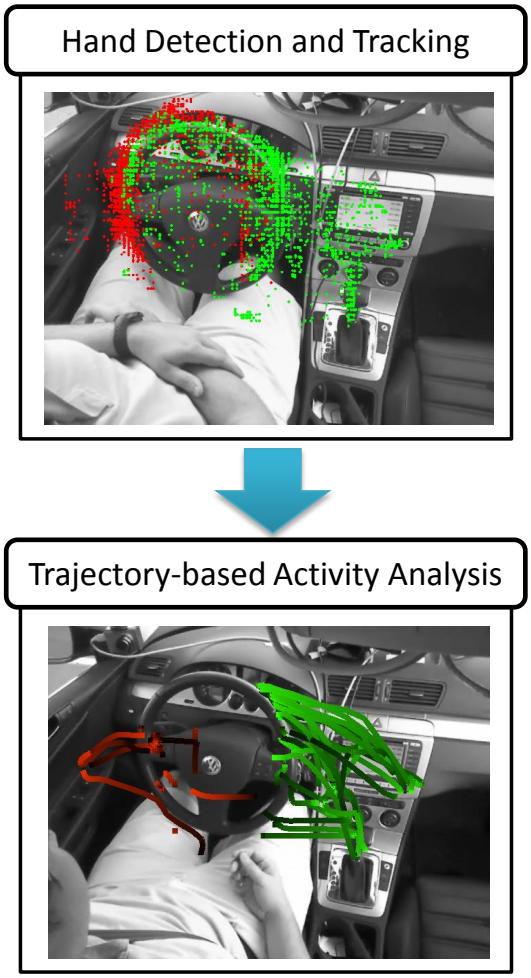


Fig. 1: Motion patterns are studied in terms of activity classification, prediction, and high-level semantics by observing hand movement in naturalistic driving settings. First, driver hands are detected and tracked in real-time in order to produce trajectories in real-time processing. The figure depicts left and right hand positions (in red and green respectively) for an entire drive. Trajectories are formed and used for several proposed driver assistance applications.

II. Hand Detection Module

In this section we specify the image pre- and post-processing, feature extraction, and training and testing rou-

†Laboratory of Intelligent and Safe Automobiles, UCSD, CA 92092, USA
{eohnbar, mtrivedi} @ ucsd.edu

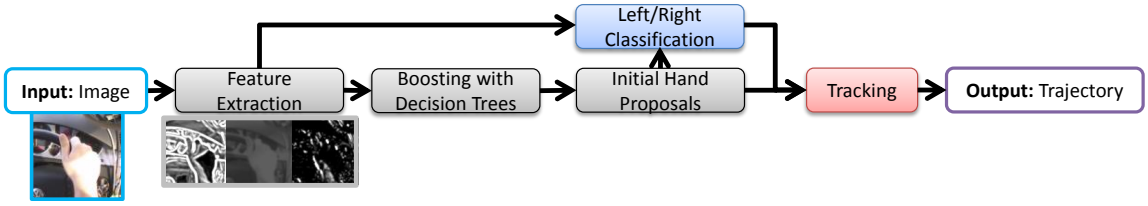


Fig. 2: The hand detection module. Hand location proposals are outputted by AdaBoost with color (LUV colorspace pixels) and gradient (normalized gradient and histogram of oriented gradients). These are classified as left or right hands, and tracking provides the hand trajectories.

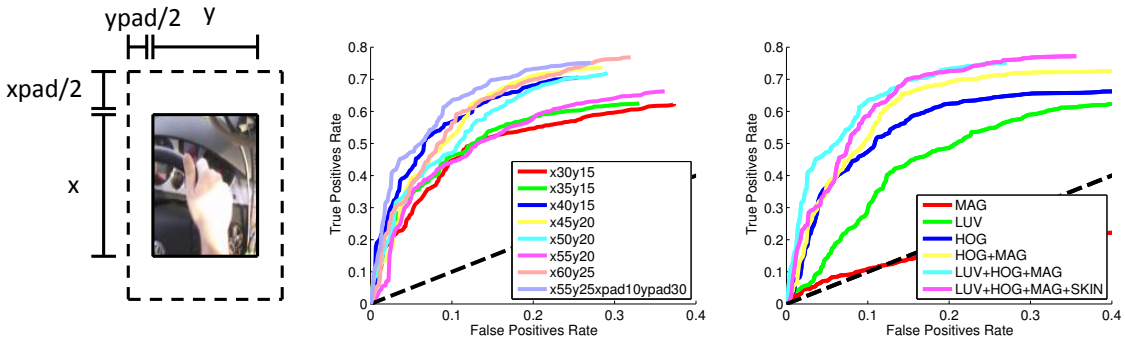


Fig. 3: Evaluation of the hand detection module using different parameters. The choice of model dimensions impact performance as only one model is trained over a variety of hand rotations and box aspect ratios. Furthermore, appropriate dimensions and padding also reduce false positives in the proximity of the hand. The impact of each of the studied features on detection performance is shown.

tines for the hand detector.

Hand detection is a challenging task, studied extensively in the vision community. In our dataset, some main challenges are common occlusion by objects and self-occlusion of the hand, deformation, and rotation (see Fig. 4. Color, edge, and texture cues are commonly used for hand detection [8], [9]). The detection scheme of aggregate channel features from [10] is employed due to the fast detection (30 frames per second on a 640×480 image) and state-of-the-art detection performance.

For evaluating the hand detection module, 922 hand instances are used for training and 1516 for testing. Color features, in particular LUV colorspace pixel values, were shown to work significantly better compared to RGB or HSV in detection. For gradient orientation features, 6 orientation bins are used. The optimal bin sizes for the HOG features computation were 2 and 4. Gradient features were processed with a normalization radius of 5. An adaboost classifier is trained in four stages, with number of trees starting at 32 and increasing by a factor of 4 in each stage. Bootstrapping is performed at each stage, with hard negatives collected and used for re-training. We experimented with additional feature channels, such as different transformations for extracting skin colored pixels using a learned skin-likelihood classifier. We found no benefit over using the simple LUV color features (Fig. 3).

As mentioned, the hand detector runs at 30 fps on a CPU, which we found crucial for analyzing hours of captured video quickly. We noticed many of the false detections occurring in the proximity of the actual hand (the arm, or multiple detections around the hand). Therefore, window size and padding had a significant effect on false positive rates (see Fig. 3). Neighboring responses were removed using non-maximum suppression with a threshold of 0.2.

Left and right hand classification: Hand proposals provided by the hand detector are given to a binary linear Support Vector Machine (SVM) [11] for left and right hand classification. The already computed gradient features are used. Color cues were not shown to be beneficial for the left/right classification. Finally, detections are tracked using a standard Kalman filter.

III. Trajectory Learning

The output of the hand detector is used as part of an activity modeling framework. Common applications with trajectory studies (e.g. surveillance) involve a set of assumptions which may not hold in our data, such as a pre-defined number of points of 'entry' and 'exit' of the moving agents for defining the activities [12]. Furthermore, trajectories that are similar semantically may contain large performance variability. For example, turning maneuvers may begin or end anywhere on the wheel, with varying velocity profiles, or with one or two hands. At times, turning may produce a very slight change in hand positions, yet we would like to recognize such events. Temporal events of no motion, which usually provide temporal segmentation information, are also

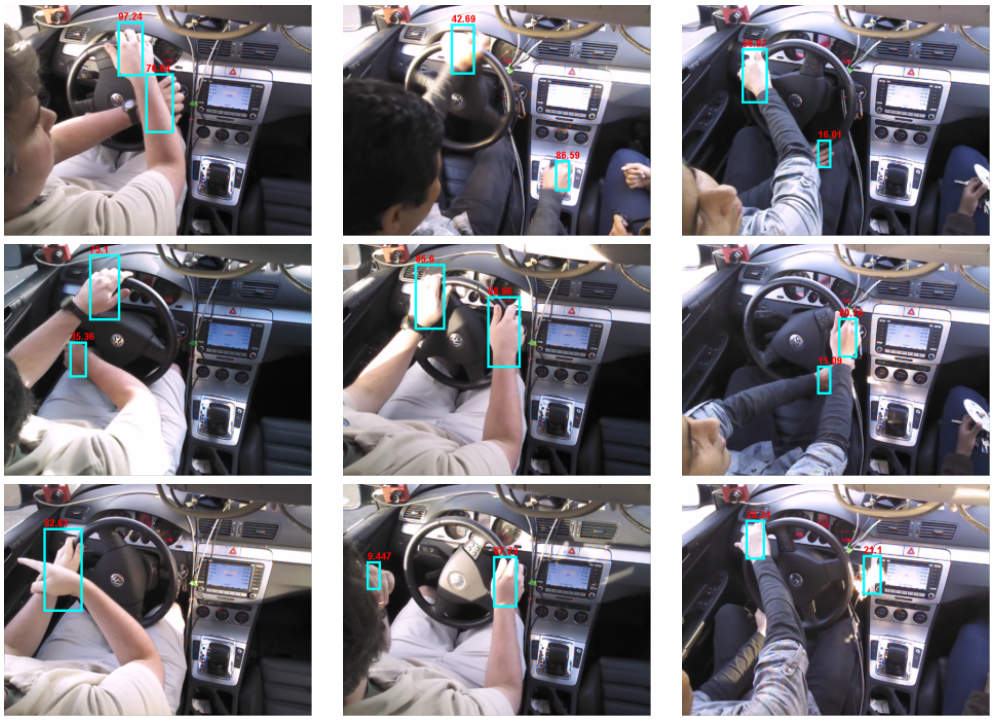


Fig. 4: Depiction of successful detection results (top two rows) and challenging settings (bottom row). The method is shown to be robust to moderate occlusion by objects in the car, self-occlusion, variation in pose and rotation. Nonetheless, false positives still occur under heavy illumination variability. These are handled by tracking.

difficult to interpret. In our domain, such 'stop' states can occur during turns, lane changes, or regular driving, and are therefore not trivially defined. In addition to distinguishing among subtle and intricate movements, gesture performance is also effected by the *preferred neutral hand* position of the driver. Because of the uniqueness of the trajectories, we turn to a careful study of both the appropriate choice of trajectory features and the temporal modeling technique.

A. Trajectory Features

The following trajectory features are studied.

Position features: A signal of the position of the hands in each frame,

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

with $j \in \{1,2,3,4\}$ so that for each dimension of position and each hand we obtain a windowed time series (for a total of $L \times 4$ sized descriptor. That is, $f_t^j \in \{x_t^{left}, y_t^{left}, x_t^{right}, y_t^{right}\}$ which are the image plane positions provided by the hand detector. L is the trajectory length.

In addition to these, trajectory shape and dynamic information can be captured in the following features.

Displacement features: Given the component displacements at time t, $\Delta f_t = f_t - f_{t-1}$, the displacement features for the trajectory are

$$V_t^j = \Delta F_t^j = (\Delta f_{t-L+1}^j, \ldots, \Delta f_t^j) \tag{2}$$

Normalized displacement features: Inspired by [13], the displacement feature vector is normalized by the sum of the magnitudes of the displacement vector

$$\bar{V}_t^j = \frac{\Delta F_t^j}{\sum_{i=t-L+1}^{t} ||\Delta f_i^j||} \tag{3}$$

Transition histogram of displacements: Proposed in [14], this histogram descriptor utilizes quantization of the displacements in V into three levels of magnitude after normalization by the maximum displacement magnitude in the trajectory. Orientation is binned into 8 sectors of the unit circle, producing a total of 24 quantization bins. Finally, a zero displacement bin is added. A transition matrix counts the frequency of occurrence from the consecutive entries in V. The final histogram descriptor is therefore of size $25 \times 25 = 625$.

Temporal pyramid of Fourier coefficients: For each dimension of F, the short Fourier transform [15] is applied and the low frequency coefficients are used. The trajectory F is recursively partitioned into levels to further capture temporal structure of the trajectory. In our experiments, we use two levels of partitioning the original trajectory, as no gains were made by further partitioning.

B. Temporal Modeling

Characterization of trajectory paths involves learning of the temporal dynamics of the hand movement. Four supervised modeling techniques are compared. An SVM classifier is studied with a linear kernel and a non-linear RBF kernel

Fig. 5: A dataset of transition reaching and retracting gestures is used for the experiments. Left hand trajectories are shown in red and right hand trajectories are shown in green. Trajectory color encodes time, with brighter being more recent in the trajectory. Shown are reaching gestures to left side rest, gear, and instrument cluster.

[11]. Both the regularization parameter C and the spread parameter γ are grid optimized. As a classical benchmark for temporal modeling, a Hidden Markov Model (HMM) learned using the Baum-Welch method and expectation maximization (EM) [16] is also evaluated. The available implementation of [17] is used, and the number of states is optimized over $\{1, 3, 5, 7\}$. A more recent development over the HMM was demonstrated with Conditional Random Fields (CRF). We employ the Latent-Dynamic CRF (LDCRF) [18], which provides an advantage over HMM due to discriminative training.

IV. Experimental Settings

The model and features will be evaluated in terms of three performance measures.

Activity classification: Each motion pattern is manually annotated with a starting frame and an end frame, interpolated to be the same size (a 20-dimensional vector), and classified into a pre-defined set of activities. The purpose of these experiments is to compare the performance of different features and classifiers. Cross-subject cross-validation is employed, where training and testing are done on disjoint subjects. Such cross-validation is employed in all of the tests below as well. Furthermore, we use **normalized accuracy** as the performance metric, where true positives in each class are normalized by the number of instances in the class before the final averaging. This takes care of unbalanced classes in evaluation.

Activity prediction: Assume an event annotation ending at a certain time, t_e. In prediction, we query the model δ seconds before t_e (i.e. at $t_e - \delta$) for a label given the sequence of observations $F_{t_e-\delta}$. There are two possible training procedures. In one, referred to as the **fixed model** procedure, only one model is trained over the annotated events once. That model is used for prediction at different δ values in testing. In the second procedure, referred to as the **shifted model**, a model is trained on samples shifted by δ (i.e. shifting δ involves re-training) and tested on the δ-shifted test samples. Both procedures allow for activity prediction, but the shifted model case requires the evaluation of multiple models corresponding to trajectory patterns specific to each choice of δ.

Abnormal event detection: Measuring the quality of the modeling can also be done on a semantic level. Can the models be used in order to distinguish critical events specific to our application domains? The important notion of 'abnormality' is a useful measure for evaluating the framework. It also allows for direct comparison with data-driven learning of models using unsupervised techniques. Traditionally, novelty detection is achieved by inspecting the scores provided by the temporal models. This is expressed in low log-likelihood scores for a CRF or HMM model. For the SVM models, we employ the point to hyper-plane distance as a confidence measure. SVM scores are normalized using coupling approaches [11]. In all cases, a cursor for the maximum posterior probability is thresholded in order to detect an abnormal event,

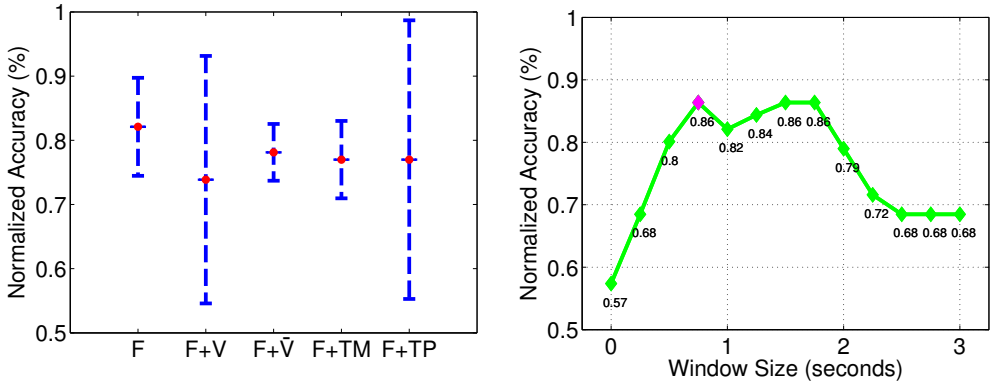
$$\max_{c \in \{1,\ldots,C\}} P(c|F) < \epsilon_{abnormal} \tag{4}$$

in a C class problem. Due to the highly complex nature of the hand trajectories, unsupervised approaches for obtaining the motion path labels may also be of interest. We also evaluate a data-driven, unsupervised trajectory analysis framework using fuzzy C-means clustering [12] and a outlier-aware K-means algorithm. In the latter case, the standard K-means iteration is performed, but at every step we use the Euclidean distance in order to discard samples that are distant from the centroid of the clusters before updating of the new centroids. The number of samples to discard is chosen according to a parameter which is fixed in each iteration. Both of the clustering algorithms contain a notion of outliers, which is essential for learning models for abnormality detection.

V. Experimental Evaluation

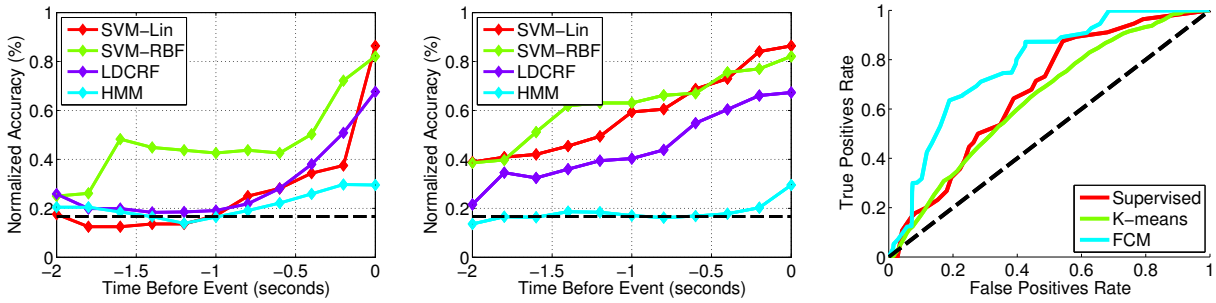
In order to evaluate the framework a video dataset composed of over an hour of driving was used. The analysis is focused on hand motion patterns which are clearly defined and are important for the study of attentiveness-reaching and retracting trajectories. Reaching motions involve hand-object interaction associated with secondary in-vehicle tasks.

Dataset: A total of 60 trajectory instances were annotated in terms of start and end, focusing on transition motions. Six classes were defined among the four regions of wheel, instrument cluster, gear shift, side rest. All trajectories must initiate or terminate on the wheel. Visualization of some of the samples is shown in Fig. 5. As the six reaching and



(a) (b)

Fig. 6: Evaluation of the trajectory features studied in activity classification. (a) Position features (**F**) are shown to work well. The abbreviations are: $\boldsymbol{V}$-displacement features, $\tilde{\boldsymbol{V}}$-normalized displacement features, **TM**-transition histogram of displacements, and **TP**-temporal pyramid of Fourier coefficients. (b) Given the annotated end of a gesture, we optimize for the temporal window size L of the time series. A 0.75 seconds window is shown to work best, and is used in the activity prediction experiments.



(a) Fixed Model (b) Shifted Model (c) Abnormal Event Detection

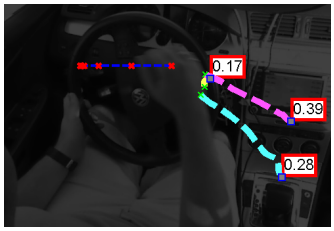
Fig. 7: Evaluation of the four modeling techniques in terms of predictive power is shown in (a) and (b). The black line depicts the random guess case. In fixed model, one model is obtained by training once using the annotated events. In shifted model, a model is learned for each δ time before an event. (c) Detecting abnormal hand activities using supervised clusters, or data-driven unsupervised clusters using K-means with outlier removal or fuzzy C-means (FCM).

retracting classes are somewhat separated in position in the image space, they are a good choice for validation and study of the different temporal features and models. Increasingly intricate motion patterns can be defined in the future. All experiments employ cross-subject cross validation, where training and testing is performed on disjoint subjects.

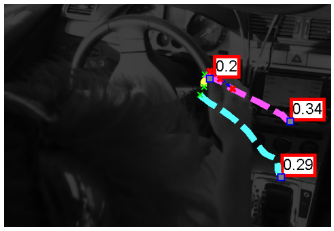
For abnormal event detection, 36 events of abnormal activity were annotated. These are events that are semantically abnormal when compared to the previous six classes of gestures which are commonly performed while driving. These include rear-mirror adjustment, driver touching the face, and driver reaching back over the shoulder to inspect and perform a reverse maneuver. Most of these involve a hand motion that is not only when abnormal compared to the six defined gesture classes, but might also be considered abnormal in certain driving scenarios (e.g. on a highway). User-specific event definition and study is left for future work.

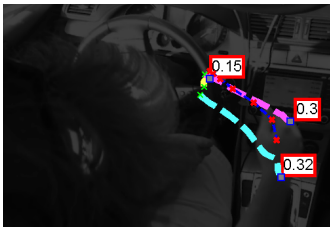
Feature analysis: On the transition motions dataset, position features alone were shown to work well out of the five types of trajectory features studied, with no clear benefit by the explicit addition of dynamic features. The analysis is shown in Fig. 6(a) in terms of the average normalized accuracy and standard deviation over the cross validation. Furthermore, given an event annotation, the optimization for the window size L to include in computation of the trajectory features is shown in Fig. 6(b). Both the position features and a window size of 0.75 seconds are employed for the remainder of the experiments. The results were produced with a linear SVM.

Temporal modeling: The four classification techniques are evaluated in terms of predictive power. As mentioned in Section IV, prediction can occur using two procedures. Overall trends are similar in both of the procedures, as shown in Fig. 7. In fixed model, where one model is trained on the annotated event end ($\delta = 0$) and evaluated at different δ values to produce predictions, an SVM with an RBF kernel is shown to work best, while a linear SVM tops for



(a) Prediction of wheel to instrument panel reaching.





(b) Prediction of wheel to gear shift reaching.

Fig. 8: Early classification of hand motion patterns. In blue is the current and previous hand trajectory (with the actual corresponding frame shown for each instance). Red crosses depict the previous hand locations in the trajectory. We plot the top three trajectories (centroids by averaging) matching to the current trajectory with the SVM probability score. Only right hand information is shown for clarity. In (a), notice how a large horizontal trajectory from the left part of the wheel is classified correctly as towards instrument cluster. In (b) note how a more difficult sample is first classified incorrectly as towards instrument cluster, but as more information becomes available the gear reaching label is correctly predicted.

classification of the gestures at $\delta = 0$. The trend is similar for the shifted model procedure (Fig. 7(b)), yet prediction rates improve overall due to the training on the shifted time series. Common ambiguous trajectories occur in reaching gestures, where a hand may reach towards the lower part of the instrument cluster or the gear shift. An example is shown in Fig. 8(b).

Abnormal event detection: The preliminary results in Fig. 7(c) shows the data-driven approach with a membership threshold using fuzzy C-means works best. In the future, unsupervised discovery of events would be essential for representing user-specific motion patterns, such as a driver's neutral hand position.

VI. Concluding Remarks

This work studied vision-based hand activity analysis. In order to tackle the intricate nature of the trajectory problem in naturalistic driving studies, multiple temporal trajectory features and classification schemes were studied in supervised settings. The framework showed promise in important applications in the context of driver safety and assistance, such as classification and prediction of gestures. The transition gestures studied and other visual-manual tasks may be correlated with head cues [19], and their integration will be studied in future work. Drive analysis techniques [20] could also benefit from hand information. Unsupervised techniques will play a key role in future work as they may allow to better capture the full range of naturalistic hand motion patterns.

References

[1] S. Y. Cheng, S. Park, and M. M. Trivedi, "Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 245–257, 2007.

[2] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Predicting driver maneuvers by learning holistic features," in *IEEE Intell. Veh. Symp.*, 2014.

[3] A. Doshi and M. M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," in *IEEE Conf. Intell. Transp. Syst.*, 2011.

[4] S. Klauer, F. Guo, J. Sudweeks, and T. Dingus, "An analysis of driver inattention using a case-crossover approach on 100-car data: Final report," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 334, 2010.

[5] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real-time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Trans. Intell. Transp. Syst.*, 2014.

[6] C. Tran and M. Trivedi, "3-D posture and gesture recognition for interactivity in smart spaces," *IEEE Trans. Industrial Informatics*, vol. 8, no. 1, pp. 178 –187, Feb. 2012.

[7] S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Hand gestures for intelligent tutoring systems: Dataset, techniques evaluation," in *IEEE Intl. Conf. Computer Vision Workshops*, 2013.

[8] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[9] E. Ohn-Bar and M. M. Trivedi, "In-vehicle hand activity recognition using integration of regions," in *IEEE Intell. Veh. Symp.*, 2013.

[10] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug 2014.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. and Tech.*, vol. 2, pp. 27:1–27:27, 2011.

[12] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2287–2301, 2011.

[13] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Intl. Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[14] J. Sun, W. Xiao, Y. Shuicheng, C. Loong-Fah, C. Tat-Seng, and L. Jintao, "Hierarchical spatio-temporal context modeling for action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[15] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

[16] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[17] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[18] L. P. Morency, A. Quattoni, and T. Darrel, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[19] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *Intl. Conf. on Pattern Recognition*, 2014.

[20] R. K. Satzoda and M. M. Trivedi, "Drive analysis using vehicle dynamics and vision-based lane semantics," *IEEE Trans. Intell. Transp. Syst.*, 2014.