Driver Activity Analysis for Intelligent Vehicles: Issues and Development Framework

Sangho Park
Computer Vision and Robotics Research Lab
The University of California, San Diego
Email: parks@ucsd.edu

Mohan Trivedi
Computer Vision and Robotics Research Lab
The University of California, San Diego
Email: mtrivedi@ucsd.edu

Abstract— **This paper examines the feasibility of a semantic-level driver activity analysis system. Several new considerations are made to construct the hierarchy of driver activity. Driver activity is represented and recognized at multiple levels: individual body-part pose / gesture at the low level, single body-part action at the middle level, and the driver interaction with the vehicle at the high level. Driving is represented in terms of the interactions among driver, vehicle, and surround, and driver activity is recognized by a rule-based decision tree. Our system works with a single color camera data, and it can be easily expanded to incorporate multimodal sensor data.**

I. Introduction and Research Motivation

Automobile has evolved from simple mechanical driving machine to intelligent semi-automatic vehicle; today's intelligent vehicle includes enhanced features for driver's safety and maneuver automation. Examples include automatic transmission and electronic fuel injection from vehicle's viewpoint, GPS navigation and lane detection from surround's viewpoint, and body posture detection and gaze analysis from driver's viewpoint. We understand that driving situation encompasses three entities, *vehicle*, *driver*, and *surround*. Most research focus has been on the analysis of a single entity of the three.

It is useful to view the driving from the viewpoint of interactions among the three entities as shown in Fig. 1.

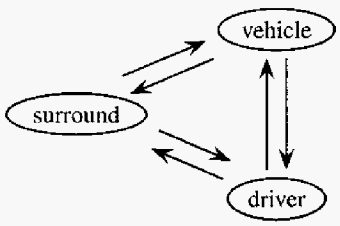


Fig. 1. driving activity analysis domain

The driver-vehicle interaction includes driver's action to the vehicle and the vehicle's reaction to the driver. It includes steering wheel operation, gear shifting, radio panel operation, etc. as driver's action, and meters' display, cockpit temperature/humidity, windshield visibility etc. as the reaction of the vehicle.

The driver-surround interaction includes the influence of surround on driver and the driver's reaction to the surround. Influence of surround includes weather conditions (i.e., shiny, cloudy, rainy, and foggy day), time-zones such as daytime/nighttime, outside temperature, etc. The driver's reaction includes attention deployment, attention distraction, action of controlling instuments, etc.

The surround-vehicle interaction includes influence of surround on the vehicle and the vehicle's reaction. The influence of surround includes road conditions (dry, slippery, bumpy, etc.), road slope (even/level, upcliming, downcliming, straight/curvy, etc.), and GPS information change. The vehicle's reaction includes translational motion, vibration, gas mileage changes, etc.

In this paper, we present a vision-based driver activity analysis system that models / represents driver's actions in terms of interactions with the vehicle and with the surround. Our approach is distinct from previous approaches, most of which aim at analyzing single body parts such as pose estimation [11], gaze detection, or affect analysis based on facial expression [4]. We aim at developing a high-level recognition scheme for driver's *activity* and semantic representation of what is going on in the driving situation.

II. Driver Activities and Interactions

Two issues are involved in driver activity analysis: *context dependency* and *resolution in activity recognition*. Firstly, in order to achieve meaningful analysis of the driver behavior, we need to consider the context in given driving situation. Driver's seemingly similar action of *gaze left* may have different connotation in sight seeing context and in passing-by context. The surround factor needs to be integrated in the recognition process. Secondly, resolution in activity recognition can be classified into gross-level recognition and detailed-level recognition. Gross-level activity may involve large-scale arm motion, torso motion, and head motion, while detailed-level activity may involve gaze change and facial expression. Detailed-level activity will require high-resolution input image frames.

With respect to safety in driving, it is important to distinguish driving patterns of intended driving vs. distracted driving [12], [9]. Although driving pattern varies across drivers, it will be possible to distinguish the two patterns, if we can build learning mechanisms for specific drivers.

0-7803-8961-1/05/$20.00 ©2005 IEEE.

DRIVER ACTIVITY HIERARCHY:

$$\begin{pmatrix} INTERACTION \\ ACTION \\ GESTURE \\ POSE \end{pmatrix}$$

Driver activity:
interaction = agent's cause action + target's effective change
action = head gesture + torso gesture + arm/leg gesture
Head-*gesture*:
associated with focus of attention
serves as a precursor of future action
Torso-*gesture*:
constrains possible configuration of body-part *poses*
associated with specific interactions
Arm-/leg-*gesture*:
constitutes action-units characterized by trajectory of arm/leg.
Body-part *pose*:
instantaneous configuration of the body part in space
at each frame.

Fig. 2. Driver activity hierarchy

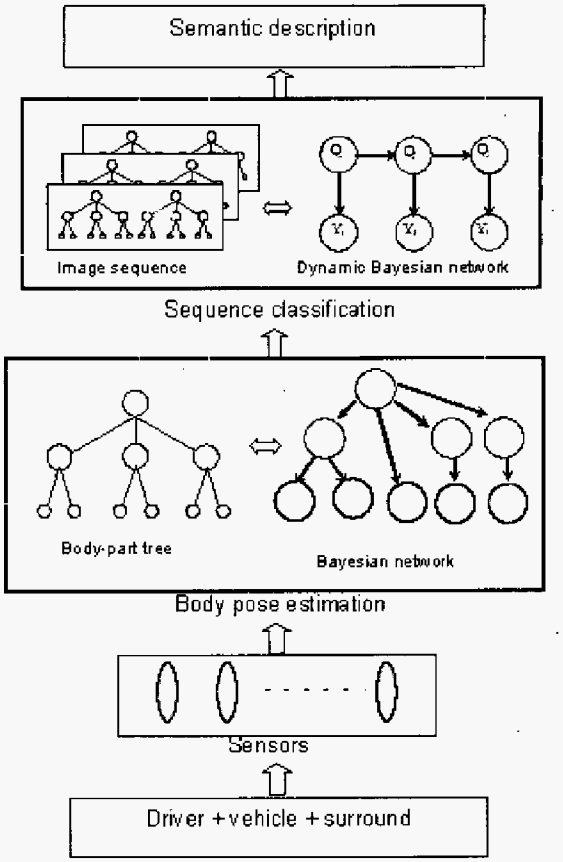


Fig. 3. System diagram for proposed driver activity analysis system

III. Driver Activity Hierarchy and System Overview

The representation of driver activity is based on the notion of hierarchy [7]; driver-vehicle *interaction* is a combination of driver's causal action and vehicle's effective change of state, and the driver's *action* is made up of multiple body-part *gestures* such as head motion, torso motion and arm/leg motion. Each body-part gesture is an *elementary event* of motion and is composed of a sequence of instantaneous *poses* at each frame. (See Fig. 2.)

The processes of the individual levels of the hierarchy are described in the system diagram in Fig. 3. The states of driver, vehicle and surround are captured by sensors. The driver's image is captured by camera and the body is segmented by background subtraction. Body parts are estimated by appearance-based partial-body model explained in Section IV-A. Body parts are represented in terms of body-part tree data structure, and the body parts' poses are estimated by a Bayesian network. Gestures (i.e. temporal evolution of poses) of the body parts are estimated by a dynamic Bayesian network. The driver's actions and interactions with the vehicle is represented in terms of semantic description. The driver's activity is recognized by a rule-based decision tree.

IV. Driver Activity Analysis

A. Partial Body Model and Cockpit Setup

Driver's body in the cockpit is captured with a single color camera. Due to spatial constraint in the cockpit area, the driver's hip and thigh are almost fixed on the seat. Drive's head position is also limited around the head rest especially while the driver starts the automobile. We build an initial occupancy map that specifies the plausible locations of the head, thigh, and torso/arms of driver in the cockpit as shown in Figs. 4 and 5. The probable locations of the body parts are represented in terms of 2D Gaussians. We also define the regions of interest (ROIs) for important vehicle equipments such as *steering wheel* (SW), *transmission lever* (TL), and *instrument panel* (IP). Window areas are also registered for background subtraction of driver's head / shoulder region.

The background subtraction gives a foreground map for the driver. Color distribution of the foreground pixels are estimated by a Gaussian mixture model (GMM) [8]. The GMM parameters are estimated by expectation-maximization (EM) learning algorithm using the initial frame data [3]. The foreground pixels of the rest of the frames in the sequence are classified with respect to the learned GMM. The pixels are grouped into blobs according to the color similarity. The foreground blobs of the initial frame are mapped to the occupancy map. Individual blob belongs to one of the body parts in our partial body model. Body part tracking is achieved by our multitarget-multiassociation tracking algorithm [8].

We first detect the head of the driver as follows; a deformable ellipse template is fitted to the silhouette of the foreground head (ellipse D in Fig. 5(b)). The reasonable range of the ellipse sizes is predefined. The ellipse-fitted head region may contain several Gaussian components that represent hair color, face skin color, etc. Appropriate Gaussians that represent

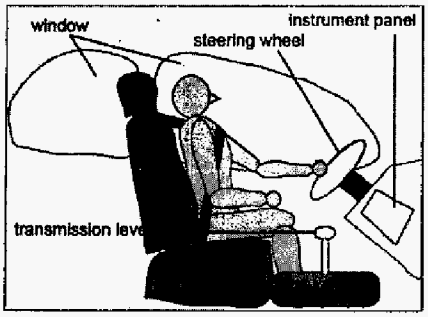
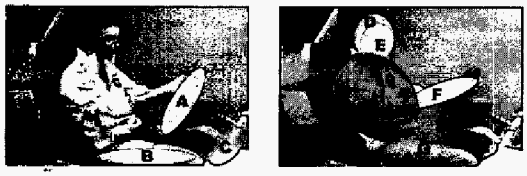


Fig. 4. Schematic cockpit setup for driver activity analysis. Region of Interests (ROI) are defined for steering wheel (SW), transmission lever (TL), and instrument panel (IP). Windows are modeled for background subtraction for driver's upper body.



(a) (b)

Fig. 5. (a)Target regions of interest (ROI) in cockpit; A: steering wheel, B: transmission lever, C: instrument panel, (b) Partial-body model represented in terms of ellipses; D: head ellipse, E: face ellipse, F: arms/hands ellipse, G: lower-body ellipse, and H: torso ellipse.

face skin color are selected from the multiple Gaussians in the head region. The skin color is initially roughly specified with threshold values in YUV color space [8]. Selecting the Gaussians trained from the actual driver image can update our initial guess of the skin color. This process corresponds to coarse-to-fine tuning of skin color distribution of the specific driver and it enables us to segment face from the head (See ellipse E in Fig. 5(b)). Other skin parts such as hands (ellipse F in Fig. 5(b)) are segmented from the foreground map by applying the selected Gaussian distributions.

We segment the lower body by selecting blobs that fall within the lower body ROI (ellipse G in Fig. 5(b)). Torso part corresponds to the rest of the blobs after removing head part and lower-body part (See ellipse H in Fig. 5(b)).

We represent individual body parts in terms of ellipses and convex hulls [6]. Kalman filters are used to reliably update the parameters along the video sequence.

B. Activity Representation

In this paper, we concentrate on examplar activities in driving behavior. The list of tested actions in this paper is summarized in Table I. The individual activity is depicted in the schematic activity patterns in Fig. 6.

Individual body parts' gesture patterns involve head motion, torso motion, and/or arm motion. Head motion includes *turning head left, turning head right, left-gaze without turn, forward-gaze without turn*, and *right-gaze without turn*. Torso motion includes *moving forward, moving backward, turning*

actions	description
drive-forward	stay right-head view and hold SW without motion
turn-left	turn head left and turn SW
turn-right	turn head right and turn SW
backup	turn head back and turn SW
touch-radio	stretch arm and touch instrument panel
shift-gear	lower arm and move lever

TABLE I

LIST OF TESTED ACTIONS. SW DENOTES *steering wheel*.

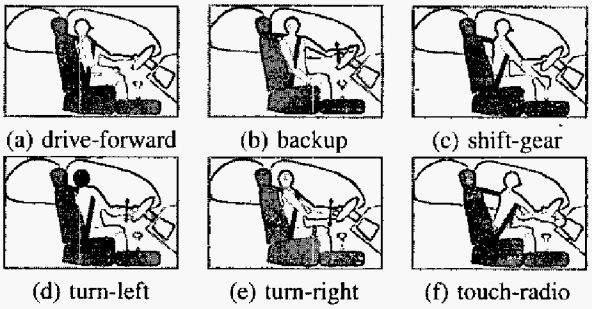


Fig. 6. Schematic activity patterns considered in the development of driver activity analysis system.

right, and *turning left*. Arm motion includes *moving up, moving down, stretching, withdrawing*, and *rotating*. Arm is the major actuator to control steering wheel, instrument panel, and transmission lever.

Pose estimation is performed by a Bayesian network that efficiently handles kinematic constraints of the body parts in the partial-body model.

Gesture estimation is performed by a dynamic Bayesian network equivalent of hidden Markov models.

The *ACTION* in the hierarchy plays key role in understanding driver activity at semantic level. It forms a semantically meaningful *event*. We represent an action in Fig. 2 in terms of *operation triplet* developed in our previous study [7] as shown in Fig. 7. We will use the terms *arm* and *hand* interchangably.

We conceptualize human actions in terms of an *operation triplet* defined as *triplet* = <*agent*−*motion*−*target*> according to the theory of 'verb argument structure' in linguistics [10]. The argument structure of a verb allows us to predict the relationship between the syntactic arguments of a verb and their role in the underlying lexical semantics of the verb. The *operation triplet* represents the goal-oriented motion of an agent (i.e., a body part) directed toward an optional target. The *agent* set contains body parts: 'head', 'torso', and 'arm'. The *motion* set contains basic 'action-atoms' as vocabulary for possible motion of the body parts: 'stay', 'rotate', 'move forward', 'move backward', 'raise', 'lower', 'stretch' and 'withdraw' . The *target* set contains important ROIs in the cockpit as vocabulary for possible target of the motion: 'steering wheel', 'transmission lever', and 'instrument panel'.

Set notation for human action:
The universe set of human action: U
U = { action | action = <agent−motion−target> }
agent set: S
S = { s_i | s_i = various body parts as agent term}
= {*head, torso, arm*}
motion set: V
V = { v_j | v_j = movement of the body part}
= {*stay, rotate left/right, move forward/backward, raise, lower, stretch, withdraw*}
target set: O
O = { o_k | o_k = cockpit elements in the schematic cockpit}
= {*steering wheel, instrument panel, transmission lever*}

Fig. 7. Driver action is represented in terms of 'operation triplet' and corresponding vocabulary sets.

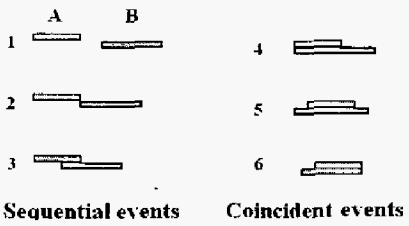


Fig. 8. Temporal relations between two events A and B: sequential (1: *before*, 2: *meet*, 3: *overlap*) and coincident (4: *start*, 5: *during*, 6: *finish*.)

C. Driver Activity Recognition

The task of activity understanding is equivalent to the task of transforming a video sequence to a semantic description using the *operation triplets* filled with the appropriate vocabulary terms in Fig. 7. The transformation rules are determined by domain-specific knowledge about driving actions. For example, the action of *touching radio* involves arm motion toward the instrument panel (IP), and is represented by <arm-stretch-IP>, while the action of *leaning forward* is represented by <torso-move-forward-null>. The *null* target indicates that no actual target is involved in the action.

Multiple events may be involved in a specific action / interaction depending on the complexity of the corresponding activity, because multiple body parts may be involved in the activity. For example, the action of *shifting gear* may involve <arm-lower-TL> followed by <arm-stretch-TL> in sequential order. The action of *touching radio* will involve coincident *operation triplets*: <arm-stretch-IP> and <head-stationary-null> with *right-view* head pose.

We adopt Allen's interval temporal logic [1] to represent the sequential and coincident relations between two events in the temporal domain (i.e., *before, meet, overlap, start, during*, and *finish* etc.) as shown in Fig. 8. Two events are called *sequential* if one follows the other within some time period. Two events are called *coincident* if the two events overlap more than half portion of each other [7].

A driver activity is described in terms of sequential / coincident relations of multiple *action* events, and the activity is recognized by a rule-based decision tree classifier. The decision tree includes domain knowledge regarding constraints in kinematically possible configurations of multiple body parts. For example, *touching radio* action may not involve *rear-view* or *left-view* head pose from the camera viewpoint, because the location of the instrument panel (IP) requires the driver to see the IP with *right-view* or *front-view* head pose. In contrast, *backing up* action may involve all kinds of head pose, since the drive may turn the head left / right or see through the rear-view mirror without turning the head while backing up. These rules are systematically organized in a decision tree, and the recognition of driver activity is a process of decision tree traversal. If the process reaches a leaf node of the decision tree, the recognition is completed.

V. System Considerations

Most behavioral analysis efforts involve simulator studies to test driver behaviors in specific situations [12]. While simulator studies permit highly controlled situations, generalization of the results to real world driving is problematic due to lack of rich context of real-world driving. We have implemented several sophisticated test beds for real-world driving in order to capture synchronized data of driving behavior and context and to analyze video and other time-based data in real time [5].

The illustrative system uses a single color camera to capture the cockpit view. The camera is firmly mounted on the passenger-side cockpit body to view the driver's upper body. Appearance-based partial body model constructed from a single camera view has ambiguity due to occlusion. It is not easy to distinguish between left and right arms.

The low-level image processing steps with only a single color camera are challenging. Alternative setup may include multiple modalities. Multiple cameras can be used for more robust segmentation of the upper body by incorporating 3D Voxel data [2], stereo range data, or thermal data [11]. Thermal camera is effective for nighttime vision. There is trade-off between robustness and cost.

Another ambiguity may be caused by sensor noise; sensor noise may make it difficult to detect the contact between a body part and a cockpit equipment. Automobile's CAN bus data [5] can be used to disambiguate the situation. CAN bus data is directly read from the automobile's electronic control unit, and it will enable us to pinpoint automobile's reaction to driver's maneuver.

VI. Experimental Results

NTSC images of drivers are captured by a Sony DCR-TRV950 handcam at frame rate of 30 frames per second. After the background statistics are collected from a set of images of the seat unoccupied, the driver is asked to perform a few driving actions summarized in Table I. Fig. 9 shows the raw and skin-segmented video frames of the drivers in different actions: *backing up, shifting gear* and *touching radio*, respectively.

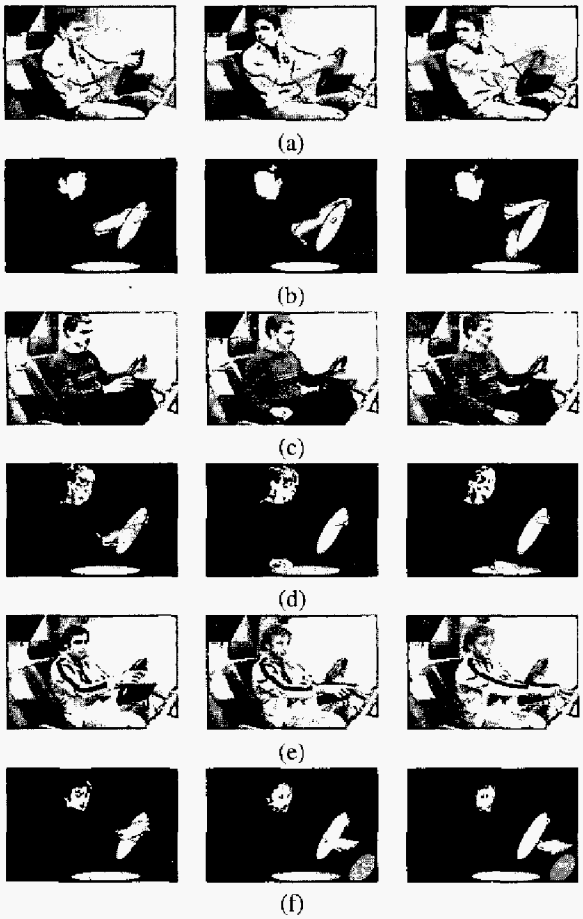


Fig. 9. Example sequences: (a) backup sequence, (c) shift-gear sequence and (e) touch-radio sequence. (b), (d), (f) are the corresponding interaction among skin blobs and cockpit ROI bolbs, respectively.

The head and hand(s) motions during the *backing-up* action in Fig. 9(a) are plotted in Fig. 10. Horizontal axes of the plots denote frame numbers spanning 1 through 160. The head orientation is plotted in Fig. 10(a). The vertical axis denotes 1: *right-view*, 2: *front-view*, 3: *left-view*, and 4: *rear-view* of the head from the camera viewpoint. The head orientation in the data changes from right-view to front-, to left-, to front-, to right-view.

The hand(s) positions in vertical and horizontal image dimensions are plotted in Fig. 10 (b) and (c), respectively. If multiple hand blobs are involved, we only consider effective blobs large enough to exceed a threshold value of predefined blob size. The position of the hand(s) is denoted by the gravity center of the pixels in the effective blobs. The vertical position denotes the distance from image top to the center of hand blob(s). The horizontal position denotes the distance from image's left border to the center of hand blob(s).

The plots show the sequential events of *head rotation* followed by *vertical hand(s) motion*. This behavior pattern is typically observed from multiple drivers in the experiments.

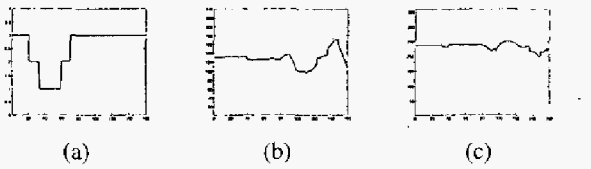


Fig. 10. Motion of individual body parts in Fig. 9(a)-(b): (a) head orientation, (b) vertical position of hand(s), and (c) horizontal position of hand(s) in terms of image coordinates. (a) vertical distance from image top and (b) horizontal position along the sequence.

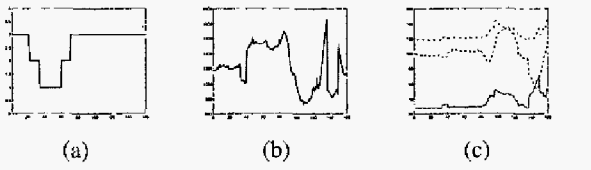


Fig. 11. Driver's *backing-up* activity represented in terms of combined actions in Fig. 9(a)-(b). (a) head orientation, (b) dispersedness of hand blob(s) represented in terms of the trace of covariance matrix from hand blob, (c) hand blob's distances from steering wheel, transmission lever, and instrument panel.

The interaction between the hand(s) and the cockpit ROIs is shown in Fig. 11. The trace of the covariance matrix of the hand pixels represents the hand blobs' spatial extension as shown in Fig. 11 (b). The temporal evolution pattern indicates that the hand(s) move around within a limited range of positions (See Fig. 10 (b) and (c).) The plots indicate that the hand(s) move around the steering wheel as shown by the short distances in Fig. 11 (c). The Eucledian distances from the hand(s) position to the ROIs are plotted in Fig. 11 (c). The bottom curve denotes the distance to the steering wheel, the middle curve to the transmission lever, and the top curve to the instrument panel, respectively.

All this low-level information is efficiently summarized and represented in terms of operation triplets in Fig. 7 by our driver activity hierarchy framework in Fig. 2.

The semantic representation of driver's activities is shown in Fig. 12 for the sequences in Fig. 9: (a) *backup* (b) *shift-gear* and (c) *touch-radio* sequences, respectively. The proposed system efficiently represents the diverse driver activities, and recognizes them correctly.

VII. Conclusion

In this paper we have shown a framework for a driver activity analysis system. Our approach is based on the hierarchy of action concepts: static pose, dynamic gesture, body-part action, and driver-vehicle interaction. We represent a driving situation in terms of multiple interactions between driver, vehicle and surround. Our framework provides a bridge to connect individual body-part tracking to the semantic level analysis of driver activity. This system can also be expanded to include other sensor modalities such as stereo image and thermal image for improving robustness. The system can also incorporate information about vehicle and surround as well as about driver for including more rich activity scenarios.

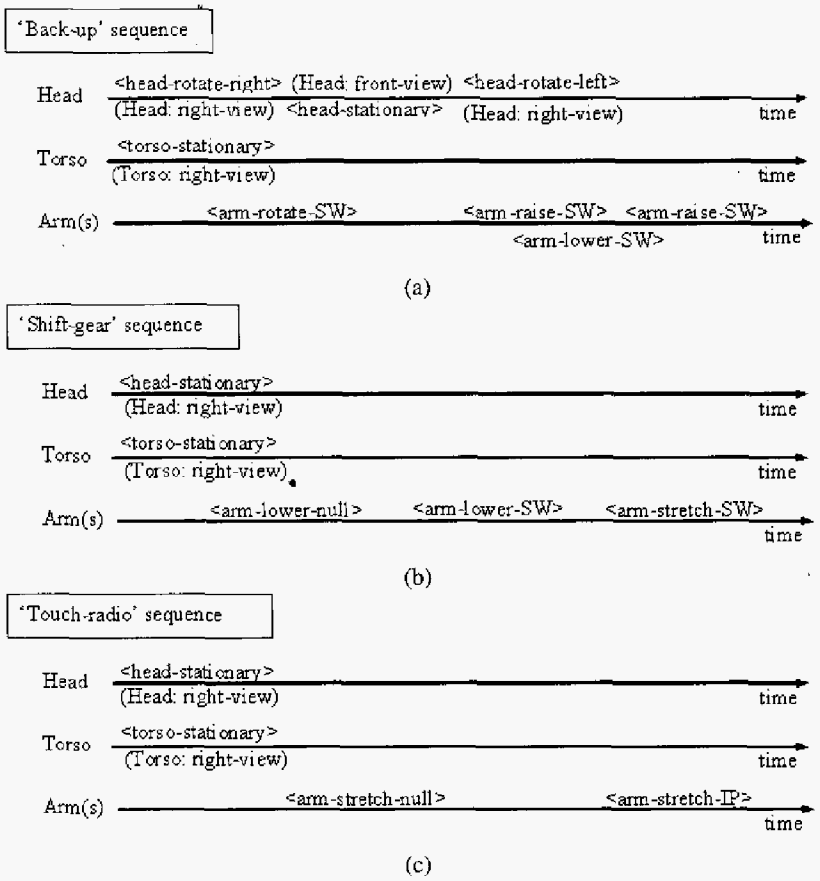


Fig. 12. Semantic representation of driver's activity in Fig. 9: (a)*backup* (b) *shift-gear* and (c) *touch-radio* sequences. Multiple body part's actions and poses are aligned along a common time line. The specific pattern of sequential and coincident action events of body parts specify the corresponding interactions between driver and the vehicle.

REFERENCES

[1] J. F. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579, 1994.

[2] S. Y. Cheng and M. M. Trivedi. Human posture estimation using voxel data for smart airbag systems: Issues and framework. In *IEEE International Symposium on Intelligent Vehicles*, pages 84–89, 2004.

[3] R. Duda, P. Hart, and E. Stork. *Pattern Classisifcation*. chapter 3, pages 84–140. Wiley, New York, 2 edition, 2001.

[4] M. T. J. McCall, S. Mallick. Real-time driver affect analysis and tele-viewing system. In *Intelligent Vehicles Symposium, Proceedings. IEEE*, pages 372 – 377, 2003.

[5] J. McCall, O. Achler, M. M. Trivedi, P. Fastrez, D. Forster, J. B. Haue, J. Hollan, and E. Boer. A collaborative approach for human-centered driver assistance systems. In *7th IEEE Conf. on Intelligent Transportation Systems*, 2004.

[6] S. Park and J. Aggarwal. A hierarchical bayesian network for event recognition of human actions and interactions. *Multimedia Systems*, pages 164–179, 2004.

[7] S. Park and J. Aggarwal. Semantic-level understanding of human actions and interactions using event hierarchy. In *IEEE Workshop on Articulated and Nonrigid Motion*, 2004.

[8] S. Park and J. K. Aggarwal. Segmentation and tracking of interacting human body parts under occlusion and shadowing. In *IEEE Workshop on Motion and Video Computing*, pages 105–111, Orlando, FL, 2002.

[9] D. Salvucci. Inferring driver intent: A case study in lane-change detection. In *Proceedings of the Human FactorsErgonomics Society*, 2004.

[10] A. Sarkar and W. Tripasai. Learning verb argument structure from minimally annotated corpora. In *Proceedings of COLING 2002*, Taipei, Taiwan, August 2002.

[11] M. M. Trivedi, S. Y. Cheng, E. M. C. Childers, and S. J. Krotosky. Posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. *IEEE Trans. Vehicular Technology*, 53(6):1698–1712, 2004.

[12] H. Yoo and P. Green. Driver behavior while following cars, trucks, and buses. Technical Report Technical Report UMTRI-99-14, The University of Michigan Transportation Research Institute, 1999.

This research is supported in part by a University of California Discovery Grant under the Digital Media Innovation Program. We are also thankful for the assistance and support of our colleagues from the UCSD Computer Vision and Robotics Research Laboratory.