ELSEVIER

Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis

Shinko Y. Cheng *, Sangho Park, Mohan M. Trivedi

University of California, San Diego, Department of Electrical and Computer Engineering, 9500 Gilman Drive MC 0434, La Jolla, CA 92093-0434, USA

Abstract

This paper presents a novel approach to recognizing driver activities using a multi-perspective (i.e., four camera views) multi-modal (i.e., thermal infrared and color) video-based system for robust and real-time tracking of important body parts. The multi-perspective characteristics of the system provides redundant trajectories of the body parts, while the multi-modal characteristics of the system provides robustness and reliability of feature detection and tracking. The combination of a deterministic activity grammar (called 'operation-triplet') and a Hidden Markov model-based classifier provides semantic-level analysis of human activity. The application context for this research is that of intelligent vehicles and driver assistance systems. Experimental results in real-world street driving demonstrate effectiveness of the proposed system.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Long-wavelength infrared; Tracking; Human–computer interaction; Gesture recognition; Template-based classification; Integration of multiple-cues

1. Introduction

Human motion and body part tracking has been an active area of research in the computer vision community in the recent past [6,9–11,21]. The recognition of person activity has many applications including human–computer interaction [24], robotics [22], and surveillance. Person activity analysis, however, is a challenging task especially in unconstrained environments due to the noise from the environments and unreliable lighting conditions. Most approaches using monocular views from color or thermal infrared cameras have limited applications. Recent advances in computer vision reflect the demands for overcoming such limitations.

In this paper, we introduce a multi-perspective (i.e., multiple camera views) multi-modal (i.e., thermal infrared and color) video-based system for robust and real-time tracking

of important body parts. The application context for this research is that of intelligent vehicles and driver assistance systems. In this context, the multi-perspective and multi-modal approach is especially important because the cockpit space in the vehicle lacks stable illumination and the driving involves dynamic change of environmental factors such as changing background and moving shadow. There are many challenges in recovering driver body movement from images of the driver in a vehicle, not least of which are caused by varying illumination during the day and between day and night, and varying appearance of different people in the driver's seat. A practical system must also have the ability of real-time (i.e., online) operation at a sufficiently fast rate.

Our overall system architecture is depicted in Fig. 1. The system receives head images from thermal and color cameras, hands images from thermal cameras, and steering wheel angle data from the CAN bus of the vehicle. Multiple feature detectors extract various features from the raw data: head orientation in terms of using the optical flow of the detected head, hand position, and velocity from tracked

---

* Corresponding author. Fax: +1 858 822 5336.
  E-mail addresses: sycheng@ucsd.edu (S.Y. Cheng), parks@ucsd.edu (S. Park), mtrivedi@ucsd.edu (M.M. Trivedi).

Head Images Hand LWIR Image Steering Wheel Angle (CAN bus)

Optical Flow

Hand Position Tracking

Grasp Analysis

Grasp Operation Triplet

Driver Activity HMM Classifier
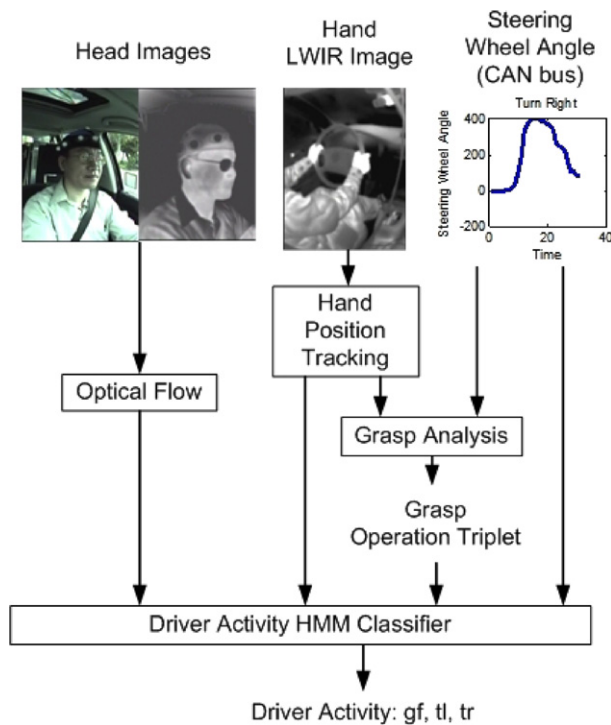
Driver Activity: gf, tl, tr

Fig. 1. The overall system architecture.

hand regions, grasp analysis with respect to the relative locations of the hands and the steering wheel, and steering wheel angle from CAN bus data. Hidden Markov model-based driver activity classifier recognizes driver activity such as 'going-forward', 'turn-left', and 'turn-right'.

The paper is organized as follows. Section 2 describes related work. Section 3 presents the tracking of hands in thermal video. Section 4 describes the tracking of the head in thermal and color video. The driver activity recognition method is presented in Section 5. The experimental studies are shown in Section 6. Finally, the concluding remarks are summarized in Section 7.

2. Related work

Annual reports on worldwide statistics of car accidents show that a considerable portion of accidents is caused by human errors such as driver's attention deficit, cognitive overload, or delay in recognizing or judging the 'dangerous' situation [2]. Driving is indeed a stressful job that involves driver's intensive cognitive processing as well as careful performance of maneuvers.

Recent research on vehicle automation has focused on development of robust automated driver assistance systems such as adaptive cruise control (ACC), crash warning system (CWS), collision avoidance system (CAS), etc. The control paradigms are acceptably established now, but more robust situation-recognition systems are required for such systems to find practical use [28]. For driver-warning systems, Donges [8] addressed that the difficulty in getting exact knowledge of driver's intention by technical

sensors would cause the warning systems to generate frequent false alarms. Therefore, a fundamental understanding of human factors such as driver's psychology and behavioral patterns is necessary in relation with the automated driving controls [4].

Information about the 3D position is recognized as an important parameter in the development of a number of safety enhancement modules. In situations where three-dimensional information about the tracks is important, multi-perspective camera-based tracking algorithms have been introduced [12]. In situations where human body parts need to be tracked in 3D space, multi-perspective voxel-based systems have shown promising results [7,20]. These studies have utilized synchronized multiple perspective color cameras as the primary source of input images.

To analyze the activity of a driver, the tracking of the key body parts (head, hands, face-gaze direction) of the driver is critical. They determine what the driver is doing, what he intends to do, as well as what he is incapable of doing at the moment (e.g., make evasive maneuvers when driving with one hand on the wheel and the other somewhere else). Some of the examples of these include, 'smart' airbags which on the basis of the 3D position of the occupant are either deployed, partially deployed or not deployed [26], driver view estimation systems based upon 3D pose of the driver's head [13], and ethnographic analysis of driving behavior [19].

This paper addresses the issue of representing and analyzing driver activity patterns by using multi-perspective and multi-modal sensing techniques. The application context is in the unified framework for integrated awareness of driving situation proposed in [27].

3. Analysis of thermal video

We propose a system that utilizes the attributes of long-wavelength infrared cameras to detect the movement of the driver's head and hands. The heat sensing attribute of the thermal infrared camera is especially appropriate for use inside a vehicle where visible illumination is constantly changing. LWIR does not exhibit problems associated with changing visible illumination since it senses emitted thermal-band electromagnetic radiation from object surfaces. A change in temperature does result in a change of the level of thermal radiation. However, this was observed to be far slower in comparison to visible illumination changes, which oftentimes occur close to the frame-rate of the camera. Fig. 2 illustrates absence of the varying illumination problem with LWIR cameras over 90 min of driving during an afternoon. The same is true from one frame to the next as it is over 90 min of constant driving as illustrated in the figure. The LWIR cameras employed in this system are DRS E3000, which use a kind of microbolometer technology that produces qualitative temperature measurements, producing images whose pixel intensity is proportional to skin's surface temperature [14,18], as compared to cameras relying on spatial photon flux differences in the sensor
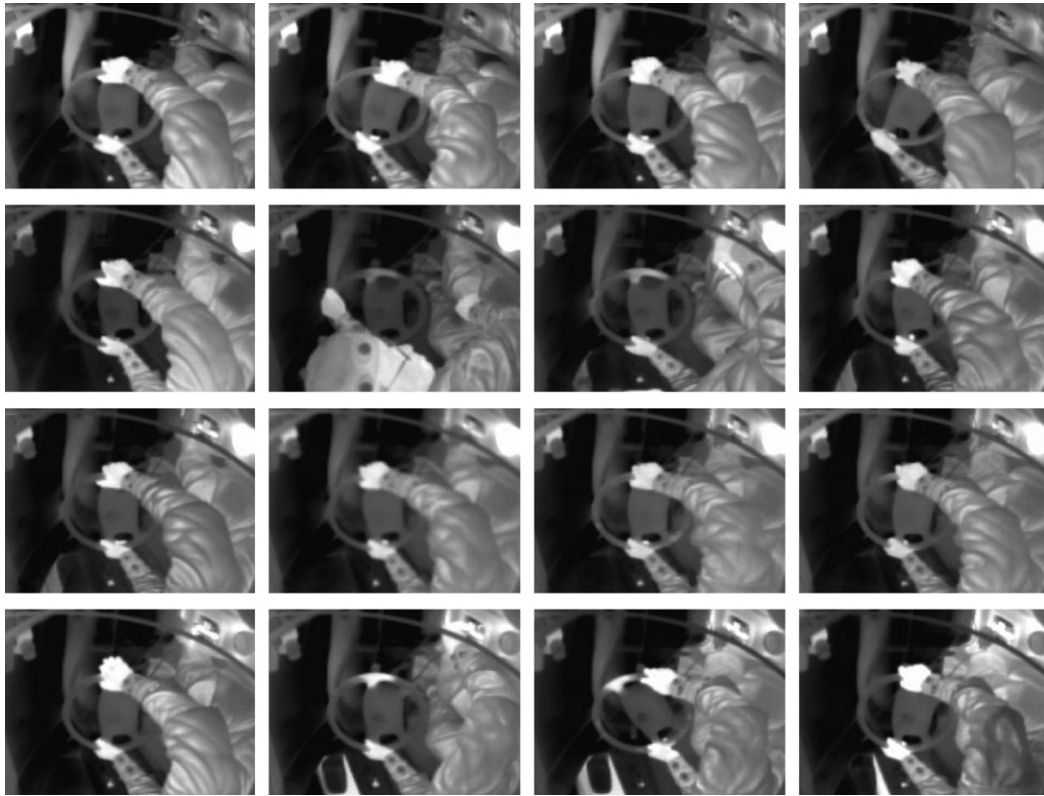
Fig. 2. These images were taken over 90 min of driving. Note that there is very little variation in the graylevels of each pixel, besides where there is movement of objects.

plane, which do not produce images with pixels proportional to temperature. This has simplifying implications on the algorithms used to extract body part attributes.

The attributes of the head are estimated with optical flow in LWIR imagery. Hand position as well as hand grasp attributes are estimated from tracking hands in LWIR imagery. This is explained in the next section.

3.1. Hand detection and tracking

As one of the cues used for recognizing driver activities, hand position in the car is the position of the hands in the LWIR image. The image location of the hands are first extracted using the Viola–Jones rapid object detector. This method uses a combination of haar-like features efficiently computed from an integral image of the LWIR image to describe regions of the image, and classification of these image regions with a boosted cascade of weak STUMP classifiers [16,29]. Fig. 3 shows examples of the 20×20 pixel positive sample thermal hand images used for training and testing. These were extracted by hand from video captured from our experimental test-bed vehicle (see Section 6). Negative examples are randomly chosen from the same video sequence everywhere except the marked hand locations. The most salient features chosen in the first three stages of the classifier cascade are shown in Fig. 4.

The Viola–Jones object detector outputs detected objects in order of likelihood. Often, only two candidate image regions are detected as hands, but occasionally more than two, one or no hands are detected when there is only
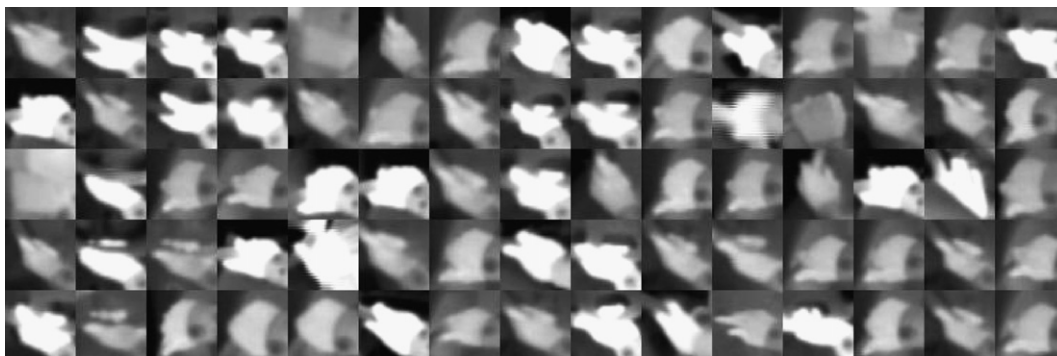

Fig. 3. Positive example LWIR images of hands of drivers. A total of 2153 examples were used.
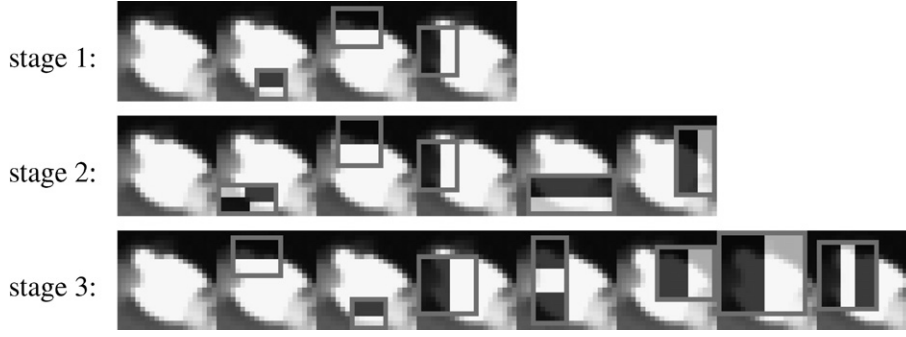
Fig. 4. Features used in the first three stages of the classifier cascade for hand detection in LWIR images.

two. To discern which of these multiple detections are truly hands and which among those are the left and right hands, we utilize a combination of kinematics information of the hand positions and their appearance. The detected hand candidates are tracked using a constant velocity Kalman Filter with Probabilistic Data Association (PDA) Filter. Multiple hand targets are maintained by producing as many tracks as necessary to accommodate all "unclaimed" measurements. The most likely targets are classified as a left or right hand by examining (1) the probability that either hand is present on either side of the steering wheel, (2) the similarity of the appearance of the target with the appearance of the last recognized left or right hand target, and (3) the longevity or confidence of the track.

The detection produces the position (x, y) and bounding-box width w of the image region detected as a hand (both left and right). This is taken as the measurement for the Kalman tracking, $\mathbf{z}_{t,m} = (x, y, w)^\top$ where $m \in \mathfrak{M}_t$ and $\mathfrak{M}_t$ is the set of measurements at time t. The state of the Kalman filter is given by the position, size, and velocity of the target hand image patch in the image. $\mathbf{x}_{t,i} = (x, y, v_x, v_y, w)^\top$.

Fundamental to the target tracking algorithm is data association, which associates candidate measurements with tracked targets. There are a number of ways to establish the correspondence between candidate measurements and targets, e.g., Global Nearest Neighbor, Probabilistic Data Association, and Multiple Hypothesis Testing methods [3,5]. Because of the sparse yet spurious errors in detection and momentary misdetections of the correct image regions, we adopt the Probabilistic Data Association method. This method collects the measurements within a specified gate, or proximity to the predicted target location, which is modulated by the estimated measurement error covariance matrix, and compares the proximity of these measurements to the predicted target location in the presence of Poisson noise. The resulting track score represents the likelihood that the measurement belongs to the track, and the measurement's likelihood to be part of the spurious background detections. These likelihoods represents the confidence of each measurement and weights them accordingly when combined to form a weighted innovation for the correction step.

Measurements that are not associated to any target during the gating process are considered new potential targets. All potential and valid targets accumulate a track confidence score E_t at time t. This score represents the proportion of times the track had a valid measurement with which to correct the estimate. This is calculated depending on whether or not a valid measurement is present to correct the estimate. In other words,

$$E_t = \begin{cases} \alpha_E E_{t-1} + (1 - \alpha_E) \cdot 1 & \exists \text{ valid measurement} \\ \alpha_E E_{t-1} + (1 - \alpha_E) \cdot 0 & \nexists \text{ valid measurement} \end{cases}$$

and α_E is the forgetting factor. Upon exceeding an empirically obtained threshold τ_E, a potential target is considered a valid target. Likewise, a target may lose track when the target does not have a valid measurement to the point when E dips below the threshold.

All potential and valid targets also maintain an adaptive appearance model $T_i \in \mathbb{R}^{M \times M}$ of the image of the hand which is updated by interpolating the detected image patch to the preset size $M \times M$ and incorporated into T_i using another first order autoregressive model (forgetting factor α_T). Similarly, appearance models are used to describe the appearance of the left and right hands, U_L and U_R. These are updated based on the appearance of the classified left and right hand targets, with a forgetting factor α_U. The appearance models U_L, and U_R are initialized to zero.

All potential and valid targets also accumulate a left hand and right hand likelihood measure. This measure consists of two quantities: (1) a target's proximity to the left or right hand's usual position in the driver's area, and (2) the similarity in appearance of the target's appearance model with the stored left and right hand appearances U_L and U_R. The first quantity is the proximity of the target to the likely locations either the left or right hand in the image which is *a priori* known. These locations are over the left and right side of steering wheel as illustrated in Fig. 5. This first quantity is modeled as a bi-variate Gaussian probability, with log-likelihood values for the left and right hand given by $l_{t,\,p} = \log P(\mathbf{x}_t | \mu_L, \Sigma_L)$ and $r_{t,p} = \log P(\mathbf{x}_t | \mu_R, \Sigma_R)$, where $\mu_L, \mu_R \in \mathbb{R}^2$. The normalized sum of squared difference is used as a measure of similarity between the appearance model of the target T_i and the left
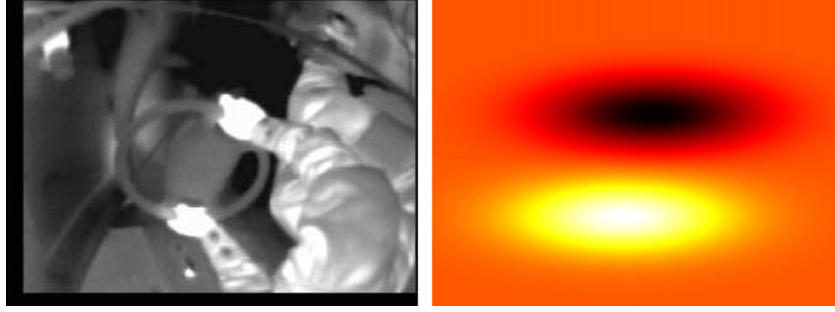
Fig. 5. Hand prior probability are illustrated here as dark and bright regions of the image.

U_{L} and right U_{R} hands. Together, the two quantities form a likelihood that the target is a left or a right hand

$$l_t = \log P(\mathbf{x}_t|\mu_{\mathrm{L}}, \Sigma_{\mathrm{L}}) \cdot (1 - \mathrm{NSSD}_{\mathrm{L}})$$
$$r_t = \log P(\mathbf{x}_t|\mu_{\mathrm{R}}, \Sigma_{\mathrm{R}}) \cdot (1 - \mathrm{NSSD}_{\mathrm{R}})$$

where $\mathrm{NNSD}_h = \|T_i - U_h\|^2/M^2$ and $M \times M$ are the dimensions of the appearance model. The amounts are accumulated with a forgetting factor α_s according to the relation

$$L_t = \alpha_s L_{t-1} + (1 - \alpha_s)l_t \qquad R_t = \alpha_s R_{t-1} + (1 - \alpha_s)r_t$$

A target with a large left-hand score relative to the right-hand score indicates that the target has hovered over the likely left hand position longer than in the likely right hand position in the image. A higher value of the left-hand score relative to all other target left-hand scores represents a higher amount of confidence and freshness of the target being and having been the left hand. Finally, among the valid targets ($E \geqslant \tau_E$), the target with the highest left-hand score and right-hand score is classified as the left and right hand, respectively.

3.2. Hand grasp analysis

Synchronous to the detection and tracking of hands from LWIR imagery, the steering wheel angle and angular velocity is recorded and combined with knowledge of hand location to determine five *operation-triplets*. Borrowed from linguistics, the operation-triplet consists of three elements: agent, motion, and target [23]. It can efficiently and completely describe the activities of the hand. The agents in this case are the left and right hands, and the target is the steering wheel or null. The null target is to describe the activity where the hand interacts with anything else besides the steering wheel. We actively detect five different motion patterns of the hands with the two targets. Each hand can

(1) grasp but not move the steering wheel,
(2) grasp and move the steering wheel counterclockwise (left),
(3) grasp and move the steering wheel clockwise (right),
(4) grasp the steering wheel loosely and allow the wheel to turn underneath it,
(5) perform other motion towards the null target.

To determine if the hand is grasping the steering wheel, an approximate model of the steering wheel is used to measure the distance between the hand and the steering wheel. This is accomplished by manually fitting an ellipse over the steering wheel in the LWIR image such that all points $\mathbf{x}$ on the steering wheel in the image is a distance of (1) away from the center of the steering wheel using the weighted 2-norm, $d(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_{\mathrm{o}})^{\top}\mathbf{S}(\mathbf{x} - \mathbf{x}_{\mathrm{o}})$ where $\mathbf{x}_{\mathrm{o}}$ is the center of the steering wheel in image coordinates. A hand detected within a certain deviation of one (1) using this model to the steering wheel using this model is considered grasping the wheel; outside a certain deviation of one (1) is considered performing other motion towards the null target.

Which of the four grasping maneuvers the hand is performing is determined by examining the coincidence of the steering wheel and hand angular velocities. For this, the angle of the position of the hands are found by finding the $\theta = a\tan((y - y_{\mathrm{o}})/(x - x_{\mathrm{o}})) - \theta_{\mathrm{o}}$ where $\mathbf{x} = (x, y)$, $\mathbf{x}_{\mathrm{o}} = (x_{\mathrm{o}}, y_{\mathrm{o}})$ and θ_{o} is the a bias applied to align the angle zero degree position to the top of the steering wheel. The angular velocity ω_h is measured by taking the difference between the last and the current angle position of the hands normalized by the duration of time passed. Then the difference is taken between ω_h and ω_{sw}. If the angular velocities between the wheel and hands exceed a threshold τ_ω, the hands, if grasping the steering wheel, are then determined to be grasping the wheel loosely and allowing the wheel

Table 1
Grasp operation-triplets

Grasp operation-triplet	Conditions
h Hand—grasp + no move—SW	$d(\mathbf{x}_h) - 1 \leqslant \tau_d$ $\|\omega_{\mathrm{sw}} - \omega_h\| \leqslant \tau_\omega$ $\theta_{\mathrm{sw}} = 0$
h Hand—grasp + turn left—SW	$d(\mathbf{x}_h) - 1 \leqslant \tau_d$ $\|\omega_{\mathrm{sw}} - \omega_h\| \leqslant \tau_\omega$ $\theta_{\mathrm{sw}} < 0$
h Hand—grasp + turn right—SW	$d(\mathbf{x}_h) - 1 \leqslant \tau_d$ $\|\omega_{\mathrm{sw}} - \omega_h\| \leqslant \tau_\omega$ $\theta_{\mathrm{sw}} > 0$
h Hand—grasp + sliding over—SW	$d(\mathbf{x}_h) - 1 \leqslant \tau_d$ $\|\omega_{\mathrm{sw}} - \omega_h\| > \tau_\omega$
h Hand—other move—null	$d(\mathbf{x}_h) - 1 > \tau_d$

SW, denotes steering wheel.

to turn underneath it. If the angular velocities are similar to within that threshold, they are then determined to be either not moving if the steering wheel angular velocity is zero, or turning clockwise or counterclockwise if the angular velocity is non-zero. Table 1 summarizes the conditions under which the various operation-triplets occur.

4. Head tracking in thermal and color videos

The thermal video provides a useful segmentation of skin areas, but the segmented thermal imagery is usually over-saturated and lacks detailed internal features. We incorporate color video and thermal video to extract additional detailed features about driver's head such as head silhouette boundary and head rotation.

We convert color video from RGB to HSV color space to separate intensity and chromaticity channels. A skin likelihood map is extracted from the hue channel of the HSV color space, and an edge map is extracted from the intensity channel by using a Canny edge detector. The Canny edge detector usually produces spurious edges in addition to genuine edges. Therefore we extract another edge map from the associated thermal video. Fig. 6 shows sub-sampled frames of color and thermal videos with associated Canny edge maps.

The detection of head position is achieved by combining the thermal and color information as follows. The candidate head position is initially detected from the thermal edge map by fitting an adaptive ellipse template using the head detection method in [15]. An ellipse of arbitrary size and tilt can be represented by a quadratic, non-parametric equation of the form:

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0, \quad b^2 - ac < 0$$

where discriminant $D = b^2 - ac$ must be negative so that the curve is an ellipse. The non-parametric form is easily converted to a parametric ellipse E in terms of the center (x_o, y_o), the length of its major and minor axes A_e, B_e, and the tilt angle θ_e: $E = (x, y, A_e, B_e, \theta_e)$.

We use a Pyramid scheme of image decimation for the edge map to achieve fast fitting of the ellipse parameter E. The fitted ellipse in the thermal image is affine-transformed to color image captured by a color camera, and used to initialize the head localization in the color image. The spurious edges in the color video are ruled out by applying a series of adaptive and elliptical ring-shaped templates that only count the edges falling between the inner and outer contours of the ring. The search range is restricted by the skin likelihood map. Fig. 7 shows the ellipse fitting process.

The head orientation estimation is performed by computing the optical flow [17] that is averaged within the head area defined by the fitted ellipse. The optical flow computation extracts appearance-based estimation of head motion which includes translation and rotation of the head. The head tracker combined with the optical flow estimator reduces the influence of head translation on the estimation of head rotation. It is observed that drivers rotate their heads frequently during driving, while the head translation occurs sporadically. This is due to the constrained structure of the cockpit and the bound posture of the driver.

5. Driver activity recognition

In this paper, we concentrate on some of exemplar activities in driving behavior: 'go-forward', 'turn-left' and 'turn-right' as depicted in Fig. 8. We recognize driver activity by analyzing the coordinated movements of a driver's head and two hands during the driver's vehicle maneuver and the glance at the outer environment. The coordinated nature of the driver activities of interest is summarized in Table 2.

A left–right Hidden Markov model (Fig. 9) is built to recognize the driver activity patterns from sequential input-feature streams. Discrete HMMs are parameterized in terms of the number of hidden states N, the number of transition links from a hidden state L, the number of observation symbols M, and probability distributions $\lambda = (\pi, A, B)$. The initial probabilities of the N states, π, is chosen from a uniform random distribution. The transition probability matrix A and the observation probability matrix B are estimated by the standard Baum-Welch

Fig. 6. Turning head in IR and color modalities and the corresponding edge maps.

Fig. 7. Head detection process: the initial ellipse on thermal edge map, ellipse search on color video's edge map, a fitted ellipse, and skin likelihood, respectively.
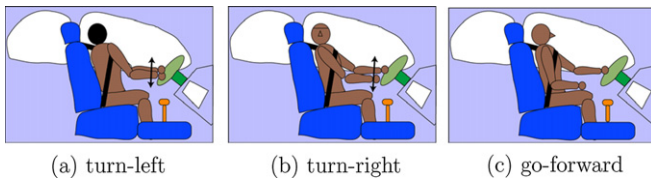

Fig. 8. Schematic activity patterns considered in the development of driver activity analysis system.

Table 2
List of tested actions

Actions	Description
Turn-left	Turn head left and turn SW
Turn-right	Turn head right and turn SW
Go-forward	Keep forward-looking head and hold SW without motion

SW denotes *steering wheel*.
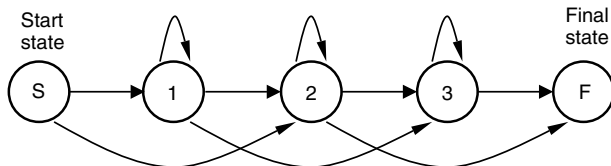
Fig. 9. A left–right HMM with three hidden states and two transition links.

Fig. 10. Process flow diagram for driver activity recognition using HMMs.

algorithm using training data [25]. The continuous input features are discretized by the Linde–Buzo–Gray vector-quantization algorithm [1]. The vector-quantization process clusters the high-dimensional input-feature space into M codewords. The input video stream data simultaneously captured from synchronized multiple IR and color cameras are used to extract continuous features such as position and velocity of head and hands converted to a series of observation symbols $1, \ldots, M$.

With the trained HMM models, new input video sequences are processed for recognizing driver activity patterns by applying a sliding moving window along time as explained below. Fig. 10 shows the process flow diagram. Sensory data streams are processed by feature extractors, and a vector quantizer converts continuous features into discrete observation symbols. The frame accumulator con-
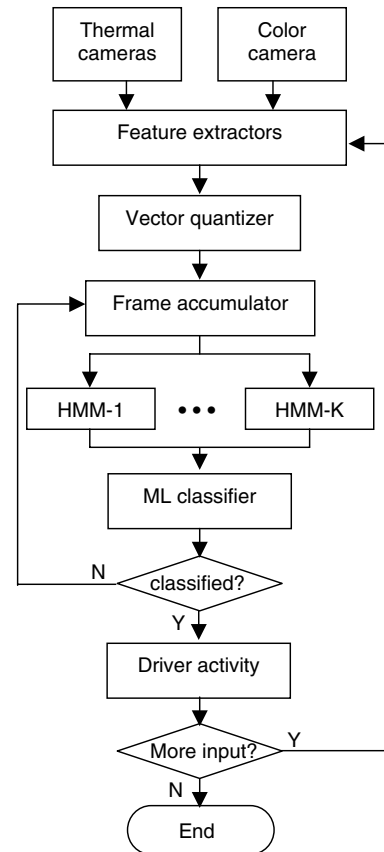
catenates the observation symbols into a stream for the input to the HMM banks. A maximum-likelihood (ML) based classifier detects driver activity patterns that have likelihood over a trained threshold. If the classification is not achieved yet, then the HMM banks receives an updated observation steam by moving a sliding window in the frame accumulator. This process iterates until there are more inputs from the sensors. Fig. 11 shows an example plot of the three HMMs' likelihood value changes as the frame accumulator's moving window slides along the observation stream. Note that the HMM bank detects the change of activity from 'go-forward' (gf) to 'turn-left' (tl) at frame 68.
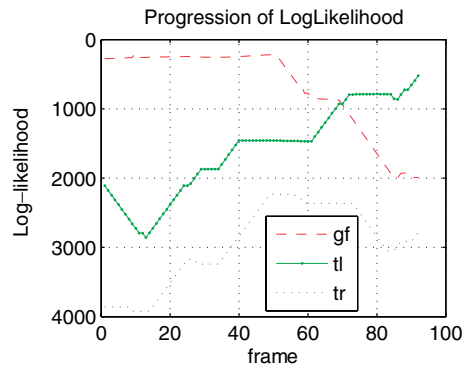
Fig. 11. Progression of log-likelihood obtained during the sliding of moving window.

6. Experimental studies

6.1. Data collection

All of the data used here was collected from a Volkswagen Passat vehicle test-bed called the for Laboratory for Intelligent and Safe Automobiles (LISA-P) shown in Fig. 13. Collecting video data from an actual vehicle is of particular significance because the data was collected under conditions close to that of a deployed vision system. The test-bed is centered on the computing, power and mounting resources in the LISA-P to allow frame-rate collection of data simultaneously from multiple sources.

To extract the movement of the driver's hands, two long-wavelength infrared microbolometer-based cameras (LWIR) were used to capture video of those events. To gain full coverage of the driver area with the 50 degrees field-of-view, they are situated above and behind the front

passenger seat is shown in Fig. 12. The cameras have a native resolution of 160×120 and are sensitive from 8 to 12 μm wavelengths. To capture the movements of the driver's head, another LWIR camera with 320×240 resolution and a color NTSC camera capture the driver's face from the passenger windshield. All of these videos are combined using a Video Combiner producing a quarter-frame image for an NTSC framegrabber.

Steering wheel angle and angular velocity as well as more than 40 other vehicle parameters are collected via a CAN bus packet translator device. A fifth camera captures the road in front of the vehicle from on top of the LISA-P to monitor its movement used primarily for video cataloging purposes. Each packet of information (video or vehicle sensor data) is accompanied by a timestamp of when the call for the data was initiated, so that data can then be time-aligned as necessary during analysis. The camera positions and views are mapped out in Fig. 13.

6.2. Hand tracking and grasp analysis

Hand detection and tracking results are shown in Fig. 14. At each time step, appearance models are maintained, and updated as image patches are accumulated from the left and right hand recognition. The row of image patches beneath the tracking illustration are the left hand, right hand, and the various target appearance models. Since all detections that are not in the gate of the other targets cause a new track to be formed, outlier targets are also tracked to disambiguate true targets with false ones. These outlier tracks tend to come and go while the true targets remain consistently tracked throughout the sequence, punctuated with moments of loss of track.
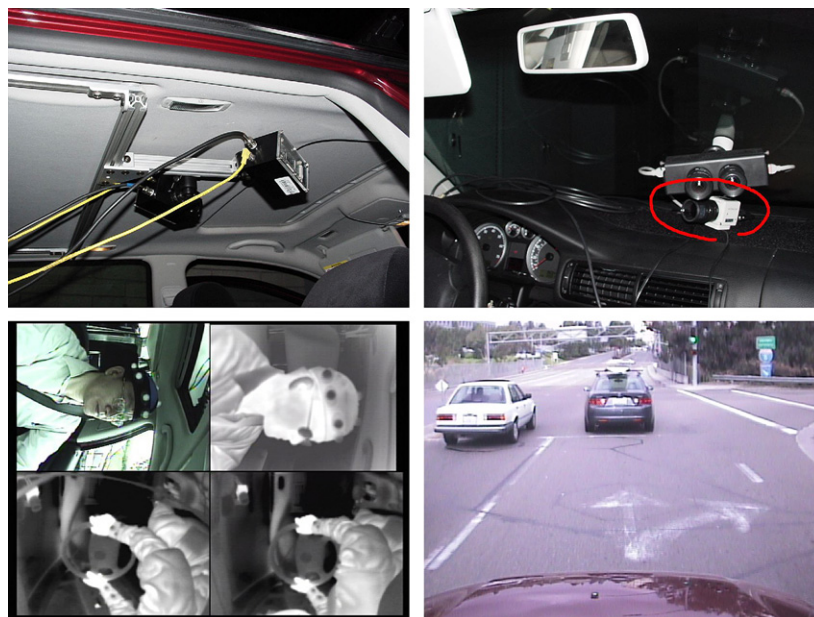

Fig. 12. Three long-wavelength infrared and one NTSC color camera are positioned around the driver cockpit area in the LISA-P test-bed to view the head and hands. A second NTSC color camera records the road ahead for event cataloging purposes.
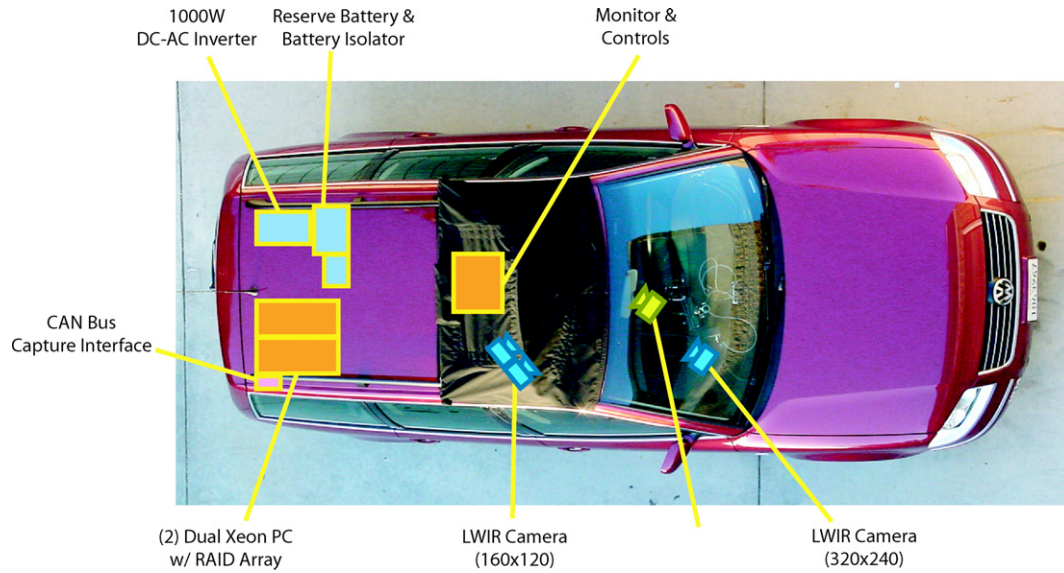
Fig. 13. Birds-eye view of the LISA-P equipped with the thermal infrared camera pair, monocular color camera, CAN bus capture card, computing, and power resources.

At the start of the algorithm, the appearance model of both hands are initialized to zero. As a result, the effect of the appearance model similarity test to discern which targets are the hands of interest have low scores on all the targets, thereby providing low influence in the decision process. Instead, only the proximity and duration of time a target lingers in the high likelihood left and right hand positions (in the image) are considered in deciding which target is the left or right hand. Then, as the appearance model is slowly updated at each time step following the target recognition, the appearance model U_L and U_R play an increasing role in deciding which targets are left and right hands. This then allows the tracking algorithm to decide targets as hands even in the rare moments when the hand targets depart from the hand's assumed usual location.

Steering wheel angle and angular velocity combined with hand position and velocity tracking and a model of the steering wheel determine which of the five grasping activities each of the driver's hands undergo. The five grasp operation-triplets are illustrated in Figs. 15 and 16.

Fig. 17 shows the histograms of the x- and y-coordinate values of the two hands during different activities. The three rows in the 3×4 grid represent 'go-forward', 'turn-left', and 'turn-right' activities, from top to bottom, respectively. The four columns represent the left hand's x- and y-coordinates, and the right hand's x- and y-coordinates, from left to right, respectively. In each cell, the abscissa and the ordinate represent frames and the corresponding coordinate values, respectively. Each cell shows the histogram of all the instances of the corresponding single activity, and the histogram value normalized by the number of the instances are represented by the brightness of color. The comparison of the three rows clearly shows distinct hands motion for these different activities. The comparison of the two leftmost and rightmost columns in a given row

shows the similar transition patterns of the left and right hands with different offsets. This analysis uncovers specific hand motion patterns in each activity.

6.3. Driving activity analysis

A set of left–right HMMs were designed with parameters $N = 3$, $L = 2$, and $M = 64$ where N is the number of hidden node states; L is the number of outgoing links from a hidden node; M is the number of codewords in the vector quantization. The parameter values were selected experimentally.

For performance evaluation purposes, we also collected steering wheel angle data from the CAN bus of the vehicle and motion-capture data using markers attached to drivers' head and wrinkles to obtain the ground truth of the vehicle status and driver's gesture, respectively. A set of independent HMMs (called 'Steering-wheel HMMs') with the same structure were used with only the steering wheel angle data to check the validity of the HMMs. A separate set of independent HMMs (called 'Motion-capture HMMs') with the same structure were used with only the motion-capture data to compare the performance of the main vision-based activity recognizer (called 'Vision-based HMMs'). All the structural parameters were the same for all the HMMs except that $M = 256$ for motion-capture data. The reason why M is bigger in motion-capture data is because its sampling rate is higher and more codewords were necessary for better representation of the data.

Training and testing data were selected by manually observing the forward-looking color camera views (Fig. 12) to determine the start and end frames for specific driver activities: go-forward (called 'gf'), turn-left (called 'tl') and turn-right (called 'tr'). This annotation procedure ensured the registration of the ground-truth segments of specific activities in terms of frame indices.
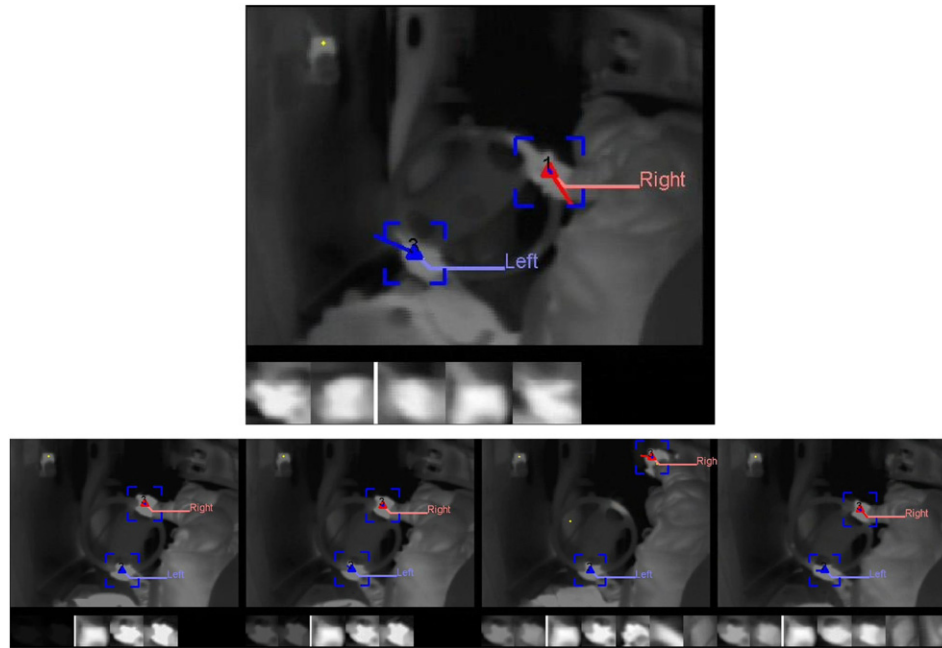
Fig. 14. Illustrated here is the progression of the first few frames of the hand tracking result. Beneath the main image are several image patches. They show the left and right hand appearance model and the track appearance models. The progression shows the left and right hand appearance model being updated over time relying on hand position prior.
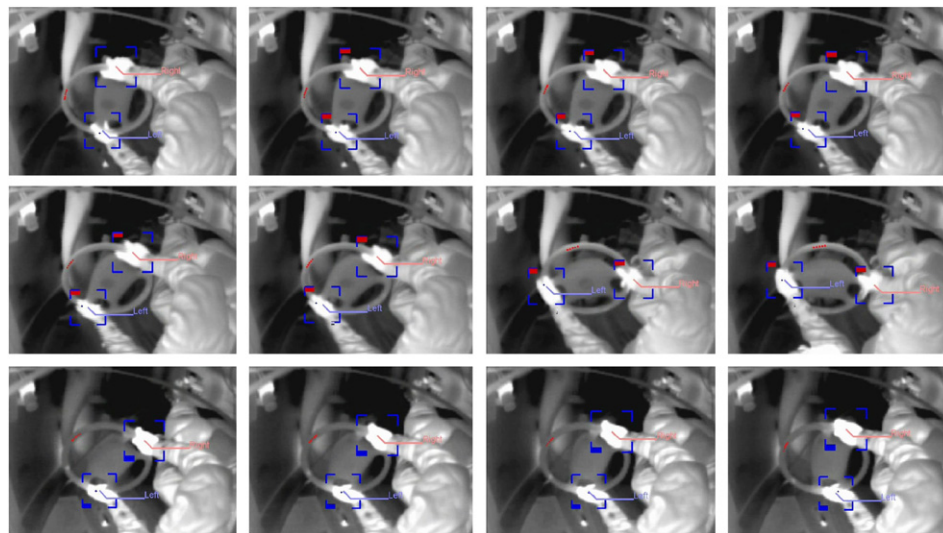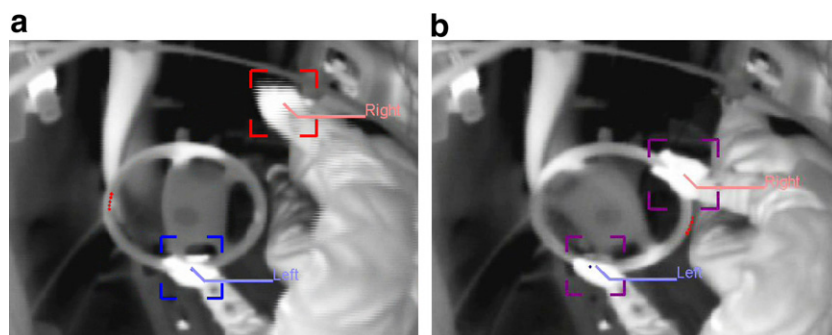

Fig. 15. This figure illustrates the two grasp types: hand on wheel turning right and left. The red and blue bricks in the bounding box represent a hand is moving the steering wheel right and left, respectively.
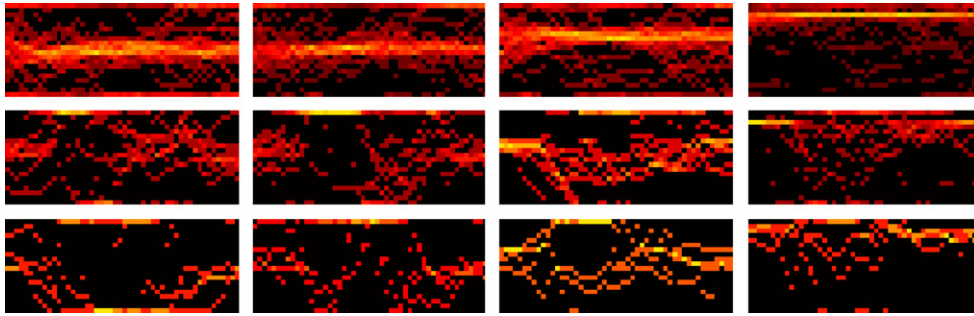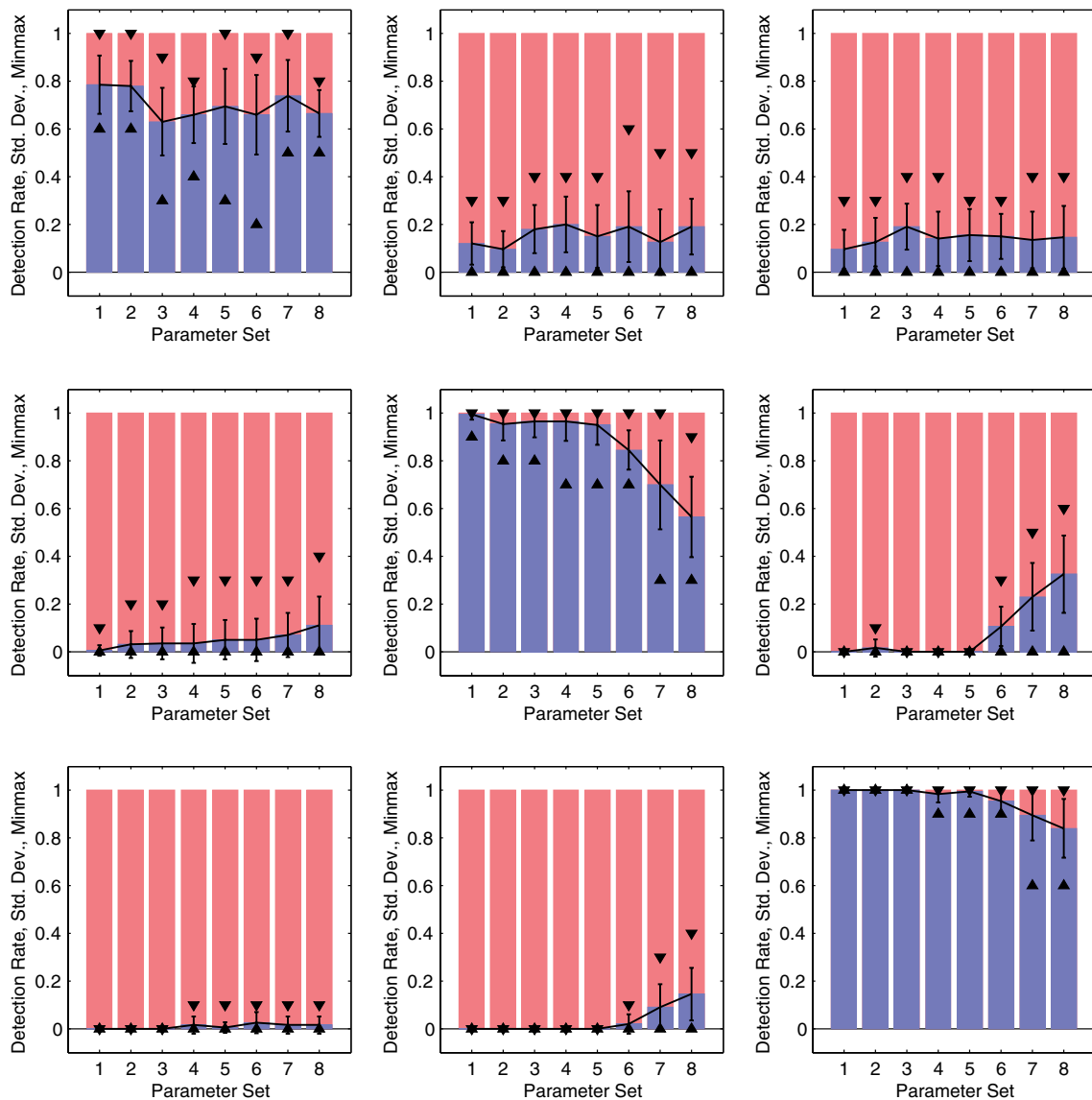
Fig. 17. 2D histogram of the hand (x,y) trajectories over time for four (4) different maneuvers.

Fig. 18. Confusion matrix of the driver activity recognition for go-forward (gf), turn-left (tl) and turn-right (tr).

Fig. 16. Illustrated here are three more grasp types: (a) left hand is grasping the wheel holding still (in blue) while right hand is away from the wheel (in red). (b) Both hands grasping but slipping over the steering wheel while the wheel straightens itself out (in purple). The grasp activity are found for one hand independently of the other.

Table 3
Tested input parameter sets for HMMs

Index	Input parameter sets	Remark
1	Steering wheel angle	Ground truth
2	Motion-capture data	Ground truth
3	Head rotation from color and thermal images + two hands' position, velocity, size + grasp analysis result	Baseline system
4	Head rotation from thermal image + left hand's position, velocity, size + grasp analysis result	Comparison
5	Left hand's position, velocity, size + grasp analysis result	Comparison
6	Left hand's position, velocity, size	Comparison
7	Grasp analysis only	Comparison
8	Head rotation from thermal image	Comparison

Total numbers of activity sequences were 28 for *turn-left*, 31 for *turn-right*, and 35 for *go-forward*. A cross-validation scheme, '10 out of X', was used to evaluate the performance of the HMM classifiers; that is, for each of the three activity classes, 10 test sequences were randomly selected for testing data and the rest of the sequences ($X - 10$) for training data, which defines one epoch of classification experiment. Total 20 epochs were made by randomly choosing different sequences for the test data. The average of the recognition results in the 20 epochs were used as the overall system-level performance measure as shown in Fig. 18. The 3×3 cells have the row indices of ground truth and the column indices of the results: *go-forward* (gf), *turn-left* (tl), and *turn-right* (tr) from top to bottom and from left to right.

We performed experiments with different combinations of input features to see what kinds of input data combination are the most effective in the activity recognition. In each cell, the bar graphs show different combinations of input parameter sets fed into the HMM classifier as shown in Table 3.

The steering wheel angle data from the CAN bus of the vehicle provides the ground truth of vehicle status during different driving activities, which is correctly trained and recognized in our HMM-based activity recognition framework. The comparison of Parameter Set 1 in each cell of the 3×3 grid provides a confusion matrix of the performance on its own for the Parameter Set 1. The motion-capture data provides another ground truth regarding the driver's body posture. The comparison of Parameter Set 2 in the nine cells provides a confusion matrix on its own for the Parameter Set 2. Other Parameter Sets can be compared in a similar way. Note that recognition performances

are moderately high for the 'go-forward' class, and very high for the 'turn-left' and 'turn-right' classes as shown in the diagonal cells. The reason of low performance for the 'go-forward' class includes the fact that 'go-forward' activity is more heterogeneous in real-driving situations, compared to the 'turn-left' or 'turn-right' that occur mainly in intersection zones (Table 4).

In a given cell, the recognition performances gradually degrade as fewer visual features are used as input. See the differences in Parameter Sets 5–8 in a cell.

The HMM-based driving activity classifier provides gross-level recognition of driving behavior such as 'gf', 'tl', 'tr', while the 'grasp analysis' provides more detailed-level information about the left and right hands' activity with respect to the steering wheel. This recognition scheme supported by the hierarchical activity grammar [23] provides a semantic-level representation and analysis of driving behavior.

7. Conclusions

In this paper, we have presented a multi-perspective, multi-modal video-based system for robust and real-time tracking of important body parts and driver activity analysis. Relying on multiple modalities and multiple perspectives per modality, the system is able to provide illumination insensitive tracking of hands and head, and fairly accurate tracking performance in noisy outdoor driving situations. A combination of deterministic activity grammar (called 'operation-triplet') and a Hidden Markov model-based classifier provides semantic-level analysis of driver activity. Experimental results in real-world street driving demonstrated effectiveness of the proposed system. The proposed framework can be applied to intelligent vehicles and driver assistance systems.

Table 4
System speed

	Time (ms)
Video and steering data capture	150
Head optical flow estimation	185△
Hand detection	56.8
Hand tracking	131△
Grasp analysis	52△
HMM classification	6.85△
Total time	582

The steps marked with (△) were implemented in Matlab.

References

[1] Introduction to Data Compression. Morgan Kaufmann Publishers, 1996.

[2] World report on road traffic injury prevention: summary. Technical report, World Health Organization, 2004.

[3] Y. Bar-Shalom, X.-R. Li, Multitarget-Multisensor Tracking: Principles and Techniques, Yaakov Bar-Shalom, Massachusetts, 1995.

[4] S. Becker, Panel discussion on introduction of intelligent vehicles into society: technical, mental and legal aspects, mental models, expectable consumer behaviour, in: IEEE Intelligent Vehicles Symposium, 1996, pp. 313–318.

[5] S. Blackman, R. Popoli, Design and Analysis of Modern Tracking Systems, Artech. House, Boston, 1999.

[6] S.Y. Cheng, S. Park, M.M. Trivedi, Multiperspective thermal IR and video arrays for 3D body tracking and driver activity analysis, in: IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum, 2005, pp. 132–139.

[7] S.Y. Cheng, M.M. Trivedi, Occupant posture modeling using voxel data: issues and framework, in: IEEE Proceedings of Symposium on Intelligent Vehicles, 2004, pp. 84–89.

[8] E. Donges, A conceptual framework for active safety in road traffic, Vehicle System Dynamics 32 (1999) 113–128.

[9] D. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1999) 82–98.

[10] D. Gavrila, L. Davis, 3D model-based tracking of humans in action: a multi-view approach, in: IEEE Proceedings of Conference on Computer Vision and Pattern Recognition, 1996, pp. 73–80.

[11] R. Hoshino, D. Arita, S. Yonemoto, R. Taniguchi, Real-Time Human Motion Analysis Based on Analysis of Silhouette Contour and Color Blob, AMDO'02, in: Proceedings of the Second International Workshop on Articulated Motion and Deformable Objects, Springer-Verlag, 2002, pp. 92–103.

[12] K. Huang, M.M. Trivedi, Video arrays for real-time tracking of person, head, and face in an intelligent room, Machine Vision and Applications 14 (2) (2003) 103–111.

[13] K. Huang, M.M. Trivedi, T. Gandhi, Driver's view and vehicle surround estimation using omnidirectional video stream, in: IEEE Proceedings of Symposium on Intelligent Vehicles, 2003, pp. 444–449.

[14] T.-L. Hwang, S. Schwarz, D.B. Rutledge, Microbolometers for infrared detection, Applied Physics Letters 34 (11) (1979) 773–776.

[15] A. Jacquin, A. Eleftheriadis, Automatic location tracking of faces and facial features in video sequences, in: Proceedings of the International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, June 1995, pp. 142–147.

[16] R. Lienhart, J. Maydt, An extended set of haar-like features for rapid object detection, in: Proceedings of the International Conference on Image Processing, vol. 1, 2002, pp. 900–903.

[17] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proc. of 7th International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.

[18] K.F. Mast, J.C. Vallet, C. Andelfinger, P.B.H. Kraus, G. Schramm, A low noise highly integrated bolometer array for absolute measurement of vuv and soft x radiation, Review of Scientific Instruments 62 (3) (1997) 744–750.

[19] J. McCall, O. Achler, M.M. Trivedi, Design of an instrumented vehicle testbed for developing human centered driver support system, in: IEEE Proceedings of Symposium on Intelligent Vehicles, 2004, pp. 483–488.

[20] I. Mikic, M.M. Trivedi, Vehicle occupant posture analysis using voxel data, in: Ninth World Congress on Intellgient Trasport Systems, October 2002.

[21] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding 81 (3) (2001) 231–268.

[22] K. Ogawara, K. Hashimoto, J. Takamatsu, K. Ikeuchi, Grasp recognition using a 3D articulated model and infrared images, in: IEEE/RSJ Proceedings of Conference on Intelligent Robots and Systems, vol. 2, 2003, pp. 27–31.

[23] S. Park, J. Aggarwal, Semantic-level understanding of human actions and interactions using event hierarchy, in: IEEE Proceedings of the Workshop on Articulated and Nonrigid Motion in conjunction with the Conference on Computer Vision and Pattern Recognition, 2004, pp. 12–21.

[24] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 677–695.

[25] L. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[26] M.M. Trivedi, S.Y. Cheng, E.C. Childers, S.J. Krotosky, Occupant posture analysis with stereo and thermal infrared video, IEEE Transactions on Vehicular Technology 53 (6) (2004) 1698–1712.

[27] M.M. Trivedi, T. Gandhi, J. McCall, Looking-in and looking-out of a vehicle: selected investigations in computer vision based enhanced vehicle safety, in: IEEE International Conference on Vehicular Electronics and Safety, Xi'an, China, 2005, pp. 29–64.

[28] A. Vahidi, A. Eskandarian, Research advances in intelligent collision avoidance and adaptive cruise control, IEEE Transactions on Intelligent Transportation Systems 4 (3) (2003) 143–153.

[29] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Proceedings of the Computer Vision and Pattern Recognition Conference, vol. 1, 2001, pp. 511–518.