2nd Joint IEEE International Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum (OTCBVS'05) in conjunction with IEEE CVPR2005, San Diego CA, USA. June, 2005.
Cheng, Park, Trivedi

Multiperspective Thermal IR and Video Arrays for 3D Body Tracking and Driver Activity Analysis

Shinko Y. Cheng, Sangho Park, Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, California, USA
{sycheng,parks,mtrivedi}@ucsd.edu

Abstract

This paper presents a multi-perspective (i.e., four camera views) multi-modal (i.e., thermal infrared and color) video based system for robust and real-time 3D tracking of important body parts.The multi-perspective characteristics of the system provides 3Dtrajectory of the body parts, while the multi-modal characteristics of the system provides robustness and reliability of feature detection and tracking. The application context for this research is that of intelligent vehicles and driver assistance systems. Experimental results demonstrate effectiveness of the proposed system.

1. Introduction

Human motion and body part tracking has been an active area of research in the computer vision community in the recent past [3, 9]. Most approaches use monocular views from color or thermal infrared cameras. In situations where three dimensional information about the tracks is important, multi-perspective camera based tracking algorithms have been introduced [4]. In situations where human body parts need to be tracked in 3D space, multi-perspective voxel based systems have shown promising results [8, 1]. These studies have utilized synchronized multiple perspective color cameras as the primary source of input images.

In this paper, we introduce a multi-perspective (i.e., four camera views) multi-modal (i.e., thermal infrared and color) video based system for robust and real-time 3D tracking of important body parts. The application context for this research is that of intelligent vehicles and driver assistance systems. Information about the 3D position is recognized as an important parameter in the development of a number of safety enhancement modules. 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 [13], Driving view estimation systems based upon 3D pose of the driver's head [5], or ethnographic analysis of driving behavior [7].

To analyze the activity of a driver, 3D tracking of the key driver body parts (head, hands, face-gaze direction) 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 (make evasive maneuvers when driving with one hand on the wheel and the other somewhere else).

The paper is organized as follows. The analysis of thermal video is presented in section 2. The analysis of color video is described in section 3. The driver activity grammar is presented in section 4. The experimental studies are shown in section 5. Finally, the concluding remarks are summarized in section 6.

2. Analysis of Thermal Video

This paper describes a system that utilizes the attributes of multi-thermal cameras to detect the 3D hand and head positions of the driver in a car. The heat sensing attribute of the thermal infrared camera is especially appropriate for use inside a vehicle where visible illumination is constantly changing. Long-wavelength infrared does not exhibit problems associated with changing visible illumination because it senses emitted electromagnetic radiation from object surfaces, which in the case of human skin is relatively constant. Human skin's tendency to emit a constant amount of radiation results in a constant intensity in infrared imagery.

Until recently, such thermal infrared cameras were very expensive and were often experimented only singly. As a consequence, detection results focused on appearance with limited 3D analysis capability. The use of multiple cameras allows localization of features in 3D space.

There have been some effort in monocular thermal infrared 3D pedestrian detection and in voxel reconstruction for hand pose analysis. To the best of the authors knowledge, there have been no prior investigations towards multi-perspective thermal infrared imagery for driver activity analysis.

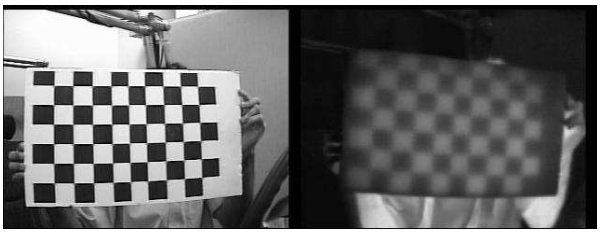


Figure 1: Calibration images.

2.1. Camera Positioning and Calibration

Four Raytheon ThermalEye AS2000 long-wavelength infrared cameras are placed over the driver cockpit with views illustrated in figure 2. Two are positioned to view the driver's hands over the steering wheel and console areas. Two other cameras are positioned to view the driver's head at about 45 degrees to one another. With these positions, activity of the driver's hands and head can be observed.

Each thermal camera is calibrated for its intrinsic and extrinsic parameters. To this end, the Camera Calibration Toolbox for Matlab is used, and a checker-board pattern warmed up with a flood lamp is used as a calibration target. Figure 1 shows how a calibration board appears after being held under the lamp for a few seconds. Following calibration, coordinates of a 3D point $\bar{\mathbf{P}} = (X, Y, Z, 1)^{\mathrm{T}}$ in the driver's space are related to the normalized camera coordinates $\hat{\mathbf{p}} = (\hat{x}, \hat{y}, 1)^{\mathrm{T}}$ and pixel image coordinates $\bar{\mathbf{p}} = (u, v, 1)^{\mathrm{T}}$ by the equation

$$\bar{\mathbf{p}} = \mathbf{KD}\hat{\mathbf{p}} = \mathbf{KD}\frac{1}{cZ}(\mathbf{R}\ \mathbf{t})\bar{\mathbf{P}}$$

where K is the 3×3 intrinsic camera parameter matrix, and $\{\mathbf{R}, \mathbf{T}\}$ as the rotation and translation constituting the extrinsic camera parameters. Radial distortion is modeled as the 3×3 matrix factor $D = diag(\lambda, \lambda, 1)$ where $\lambda = 1 + \kappa_1 d^2 + \kappa_2 d^4$ and $d^2 = \hat{x}^2 + \hat{y}^2$.

The detection of the hands of the driver begins by segmenting infrared imagery of the driver for exposed skin areas using an upper and lower intensity threshold. Noise is eliminated from the segmented images by labeling the blobs and enforcing a minimum blob area and blob perimeter. In detecting the hands, the remaining blobs at this stage in each of the two images are then associated with one another using the epipolar constraint $x'Ex = 0$, where x, x' are the points in the two images in normalized image coordinates and $E = RT_{\times}$ is the essential matrix. The points that are successfully paired up are then triangulated using the camera intrinsic and extrinsic parameters.

Skin segmentation is also the first step for head detection, after which an edge based approach is taken to find the best fit ellipse. Several ellipse templates of different dimensions and eccentricities (tilt) are matched against the downsampled edge map of the segmented image. The head image coordinate is estimated by the center of the highest scoring ellipse candidate. The head detection algorithm is described and evaluated in [13]. The hand detection algorithm is explained in further detail below.

Figure 2: Four long-wavelength infrared and four NTSC color cameras are positioned around the driver cockpit area to view the head and hands.

2.2. 2D Image Segmentation and Detection

The detection method relies on the constant temperature and emissivity of skin as perceived by the thermal camera. Aside from the camera's auto gain control or other mechanism that changes the intensity-to-temperature relationship, the intensity response from skin remains relatively constant. Using an upper and lower threshold, the skin pixels can be isolated, as shown in figure 3. For our experiments, these thresholds were $(\tau_l, \tau_u) = (210, 250)$ in a scale of 0 to 255.

In the case of hand detection, the outlying blobs are easily filtered out by enforcing a minimum blob area and blob perimeter, as mentioned earlier. In choosing a threshold, we manually classified several hundred blobs over several sequences of test video, and empirically chose a threshold appropriate to the infrared image resolution (320,240) and viewpoint (50 degrees horizontal field of view). The chosen thresholds were a lower and upper threshold of 200 and 1200 for the area, and lower and upper threshold of 3 and 9 for the compactness, which is defined as the blob area divided by blob perimeter. Blobs that remain are further filtered out by the enforcing of epipolar constraints. In the case of multiple blobs aligned with a single epipolar line, the blob with the smaller deviation from the epipolar line and satisfies the left-right consistency constraint [14]. The result is fairly reliable detection of the hands.

2.3. 3D Detection and Tracking

Following 2D isolation of the hand and head blobs, the 3D position is found by triangulation, a method described by Trucco and Verri [14] using the pre-computed camera parameters. Essentially, three vectors are formed by the camera centers and the image points, all in left camera coordinate system: $a\mathbf{p}_l$, $\mathbf{T} - bR^T\mathbf{p}_r$, and $c(p_l \times R^T\mathbf{p_r})$. Their sum results in a path about the two camera origins and the 3D point in space. The length of each ray is parameterized by (a, b, c) which we solve for using Least Squares from the the system of equations

$$a\mathbf{p}_l - bR^T\mathbf{p}_r + c(\mathbf{p}_l \times R^T\mathbf{p}_r) = T \qquad (1)$$

where $\mathbf{p}_l$ and $\mathbf{p}_r$ are the point projections in the left and right image, and $R = R_r R_l^T$ and $\mathbf{T} = \mathbf{T}_l - R^T\mathbf{T}_r$ are the relative camera orientations given by the extrinsic parameters R_r,R_l,$\mathbf{T}_l$, and $\mathbf{T}_r$. After solving for (a, b, c), the 3D point with respect to the left camera coordinate system is $a\mathbf{p}_l$ and the error residue is given by $0.5c(\mathbf{p}_l \times R^T\mathbf{p}_r)$.

The 3D points are associated across frames by using proximity to the last tracked 3D point as the criterium. New tracks are formed when point observations are not close enough to any of the last tracked 3D points (tracks). Existing tracks are eliminated when the track has no corresponding observation in up to 3 consecutive frames. This procedure does not allow us to automatically determine nor maintain left and right hand tracks. We leave that for future work. This procedure does allow us to produce valid tracks of left and right hand blobs for several hundred frames at a time, which we manually extract for use for the activity analysis explained below.

3. Analysis of Color Video

We adopt the codebook-based background subtraction method [6] to segment foreground regions of the color video frames (Fig. 6 (a) and (b)). After background subtraction, we get a foreground map for the driver. The foreground map in Fig. 6 (b) is projected vertically and horizontally to obtain projection profiles. The vertical projection (Fig. 6 (c)) is then modeled by a Gaussian mixture model (GMM) [10]. The GMM parameters are estimated

Figure 3: Skin segmentation by intensity thresholding.

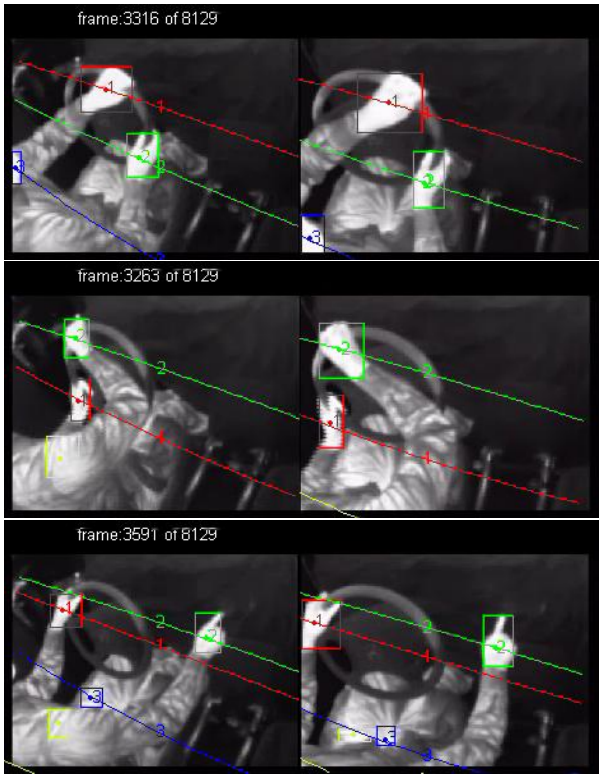


Figure 4: Illustrated here are three resulting instances of the hand detection and correspondence matching. The lines represent epipolar lines generated from points in the other image. The green (1) and red (2) boxes surrounding the left and right hand blobs survived the two stage blob filtering process: blob characteristics and epipolar constraint.

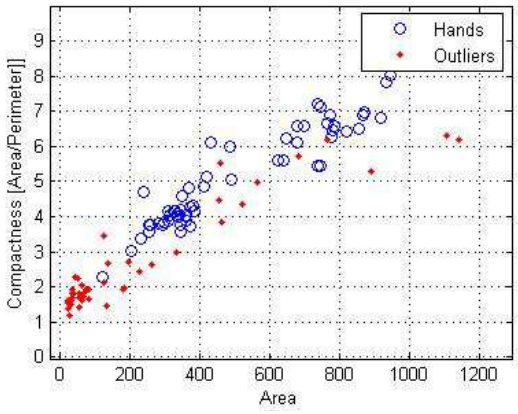
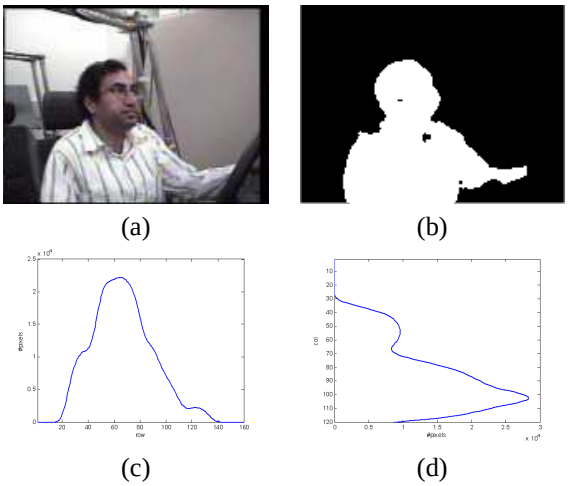
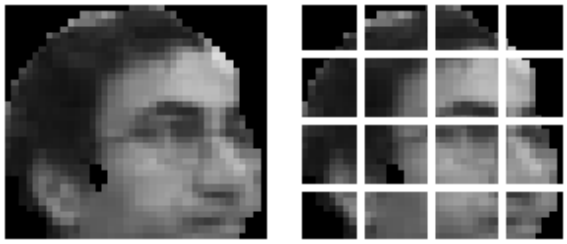


Figure 5: Plot of the compactness versus area of hand blobs and non-hand/head blobs.



(a) (b) (c) (d)

Figure 6: Background subtraction and head detection in color image. (a) Raw input image, (b) foreground map, (c) vertical projection of the foreground map, and (d) horizontal projection of the foreground map.



(a) (b)

Figure 7: Grid generation for head orientation classification

by expectation-maximization (EM) learning algorithm using the initial frame data [2]. The mode value of the projection profile provides the head position in horizontal image dimension. The horizontal projection profile (Fig. 6 (d)) provides a mode and several dips. The first dip position upward from the mode point provides the effective location between the head and torso.

3.1 Head Orientation Classification

The detected head is tracked and cropped to classify its orientation (Fig. 7.) Our head orientation classifier is view-based and does not use a specific head model or face model in order to handle a head image arbitrarily oriented on a horizontal plane.

The classifier uses moments of intensity distribution of pixels by masking a 4×4 grid on a head image as shown in Fig. 7(b). The moments-based algorithm makes the classifier independent of image size and robust against pixel loss due to imperfect ellipse fitting or imperfect background subtraction. Moments are computed for each of the 16 grid elements, respectively, using nonzero pixels in the grid. Extremely bright outlier pixels due to eyeglasses or earrings etc. are excluded from the computation by using a threshold calculated based on the overall intensity distribution of the head images in the training data. The 4×4 grid generates a feature vector X composed of 16 moments values:

$$X = [W_1, \ldots W_{16}]^T \quad (2)$$

where each W_i is computed as follows:

$$W_i = \frac{[(1/N_i)\sum_{j,k} I_i(j,k)] - M}{SD} \quad (3)$$

where $I_i(j,k)$ is the nonzero pixel intensity at pixel location (j,k) in window W_i of the grid, and N_i is the number of non-zero pixels in W_i. M and SD are the mean and the standard deviation, respectively, of the overall nonzero pixels in the cropped input head image. Classification of head orientation is performed by using a K-nearest neighbor classifier. The K-nearest neighbor classifier assigns a feature vector X to class Ω_i if the feature vector has the largest number of nearest neighbors belonging to class i among the K nearest neighbors. Discriminant function $g_k(X) = -d(X,X)k), k \in \{1,..,32\}$ is a measure of the distance between X and the feature vector X_k in the training set. The distance function $d(X, X_k)$ is selected as the L_2 norm:

$$d(X, X_k) = \| X - X_k \| = [(X - X_k)^T (X - X_k)]^{1/2} \quad (4)$$

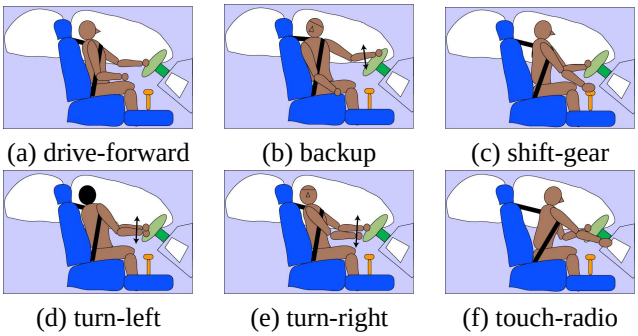
actions	description
drive-forward	stay right-head view and hold SW without motion
turn-left	turn head left and turn SW
turn-right	turn head right and turn SW
backup	turn head back and turn SW
touch-radio	stretch arm and touch instrument panel
shift-gear	lower arm and move lever

Table 1: List of tested actions. SW denotes *steering wheel*.

4. Activity Grammar

In this paper, we concentrate on examplar activities in driving behavior. The list of tested actions is summarized in Table 1. The individual activity is depicted in the schematic activity patterns in Fig. 8.

Individual body parts' gesture patterns involve head orientation, head translation, and arm motion. Head orientation includes *turning head leftward, turning head rightward, stay with leftward gaze, stay with forward gaze*, and



(a) drive-forward (b) backup (c) shift-gear

(d) turn-left (e) turn-right (f) touch-radio

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

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*}

Figure 9: Driver action is represented in terms of 'operation triplet' and corresponding vocabulary sets.

stay with rightward gaze. Head translation includes *moving forward, moving backward, moving down, leaning right*, and *leaning left*. Arm motion includes *moving up, moving down, moving leftward, moving rightward, stretching*, and *withdrawing*. Arm is the major actuator to control steering wheel, instrument panel, and transmission lever.

Pose and gesture estimation is performed by a dynamic Bayesian network equivalent of hidden Markov models.

We conceptualize the driver activities in terms of an *operation triplet* defined as *operation triplet* ≡ <*agent−motion−target*> according to the theory of 'verb argument structure' in linguistics [12] (See Fig. 9). 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'.

The task of event 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. 9. 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.

5. Experimental Studies

NTSC images of drivers are captured by 4 IR cameras and 4 color CCD cameras, all of which are calibrated. 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 1.

We performed sensitivity analyses for pixel error to 3D spatial resolution for our current camera set-up in Fig. 10. To measure the accuracy of the triangulation in real-world units, a uniform array of points is constructed in the visible space of the two front viewing cameras and the points are projected onto the image plane. The images are then subject to uniform noise to simulate a fixed amount of pixel deviation in all directions. The image points are then triangulated and the RMS error is plotted against the mean of the uniform noise process. This test is done several times for a given mean and the error is averaged. The plots shown illustrate the accuracy of the thermal feature triangulation technique up to the accuracy of the camera calibration. These assume a pair of 320 $\times$ 240 images using the 2 color and 2 infrared cameras ahead of the driver. If we measure our driver feature with nearly 10 pixels in error, we will produce on average 30mm deviation from the actual feature location in 3D space, an estimated 60mm maximum deviation. So if we can keep the error to within 5 or 10 pixels, the 3D triangulation results remain reliable.

The head and hand motions during the *touch-radio* action are plotted in Fig. 12. In Fig. 12 (a) the head positions are denoted by the red dots over each frames, which remain at the approximately the same position in 3D space. The Euclidean distances from the head and the right hand to the console, aisle, wheel, and seat are plotted in Fig. 12 (b). Whereas the head position is stable, the hand starts approaching to the console at about Frame number 100 and stays close to the console for about 40 frames. The hand

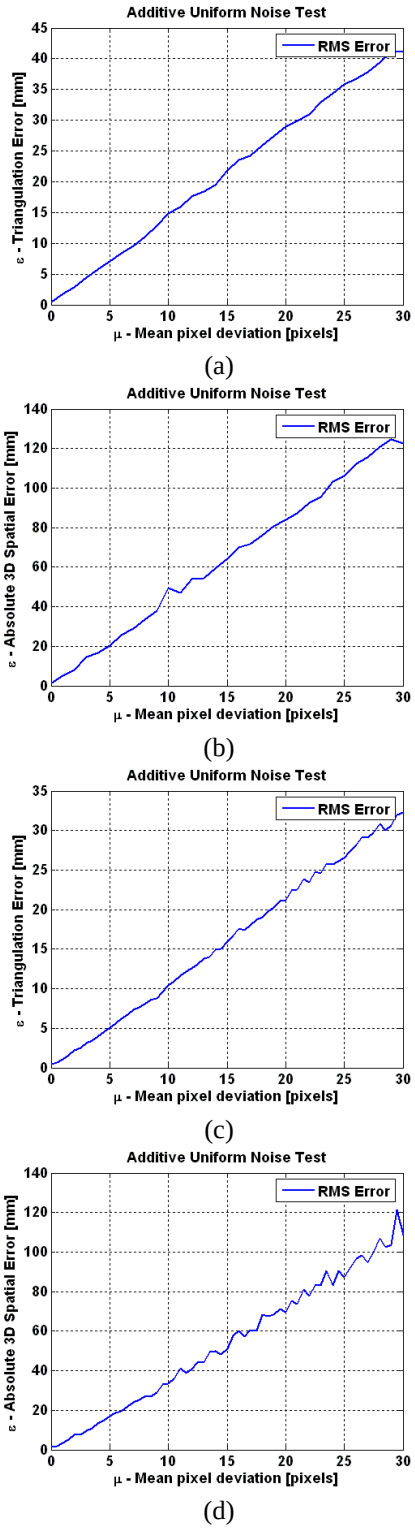


Figure 10: Sensitivity analysis for pixel error to 3D spatial resolution. (a) Uniform noise vs. triangulation error in IR camera, (b) Uniform noise vs. 3D spatial error in IR camera, (c) Uniform noise vs. triangulation error in color camera, and (d) Uniform noise vs. 3D spatial error in color camera.

motion between frame number 100 and 110 is detected by a DBN equivalent to HMM [11]. The target of the hand motion corresponds to the console according to the proximity in the plot. We set up the threshold for proximity as 150 mm. The threshold may be adjusted according to the resolution of image input.The semantic representation of the drivers activity is represented in terms of operation triplets juxtaposed along a common timeline and is shown in Fig. 14 for the sequences *touch-radio*.

The head and hand motions during the *turn-right* action are plotted in Fig. 13. The head detection and tracking results in *backing-up* action viewed from the camera pose in Fig. 6 is plotted in Fig. 15. The corresponding results of head orientation classification are plotted in Fig. 11. The class labels of head orientations viewed from camera start from 1: 'rear view' of the head to 8: 'rear/right-profile view' (i.e., similar to viewing from camera rotating clockwise around the head at 45 degrees each). Given the known camera pose in Fig. 6, it is straightforward to convert the head orientation from the camera-centered coordinate system to the driver-centered coordinate system. For example, the 'front/right-profile view' in camera-centered coordinate system in Fig. 6 corresponds to 'front view' in the driver-centered coordinate system in which the driver looks ahead toward the dashboard. The plots show the systematic change of head orientation in *backing-up* action.

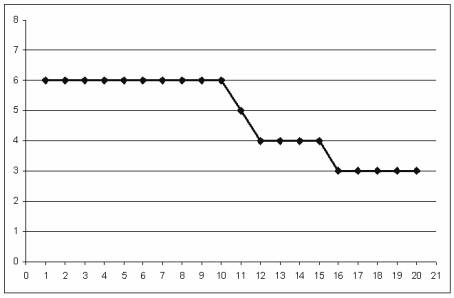
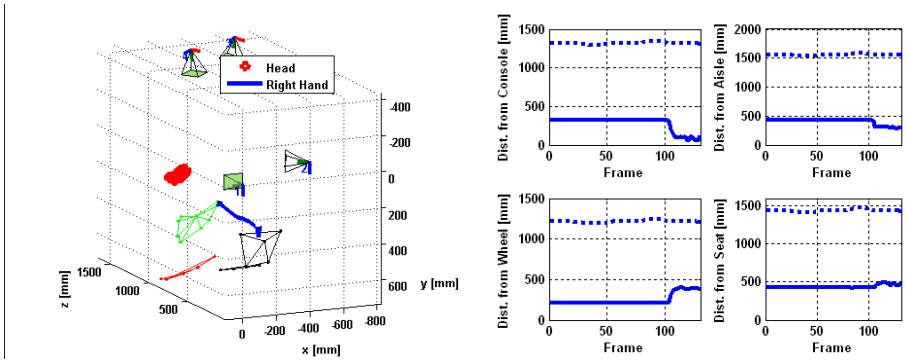


Figure 11: Classification of head orientations in Fig. 15: Frame number vs. Orientation class labels. See the text for details.

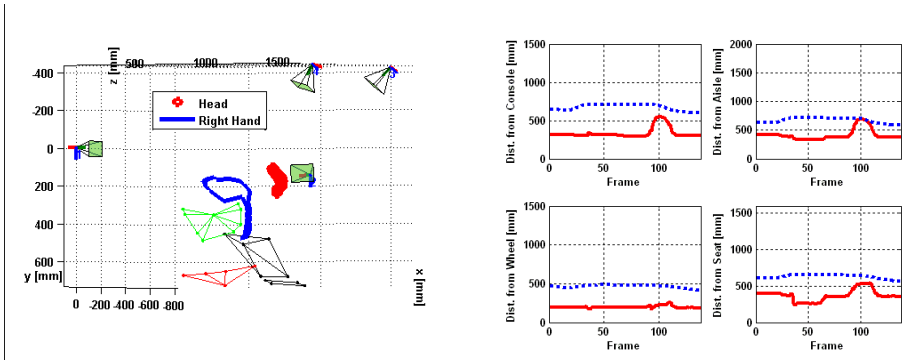
6. Concluding Remarks

In this paper, we have presented a multi-perspective, multi-modal video based system for robust and real-time 3D tracking of important body parts. Relying on multiple modalities and multiple perspectives per modality, the system is able to provide illumination insensitive tracking of hands and fairly accurate 3D tracking performance in noisy environments (30mm average spatial error with 10 pixels deviations). The richness of 3D based feature detection and tracking has showed the effectiveness of the system as an activity analyzer of complicated tracking tasks of the driver.



(a) (b)

Figure 12: 3D trajectory (a) and Euclidean distances (b) of the right hand (solid lines) and the head (dotted lines) to cockpit objects in *touch-radio* action. The irregular grids in green, red, and black denote the 3D positions of the steering wheel, the seat, and the console, respectively.



(a) (b)

Figure 13: 3D trajectory (a) and Euclidean distances (b) of the right hand (solid lines) and the head (dotted lines) to cockpit objects in *turn-right* action.

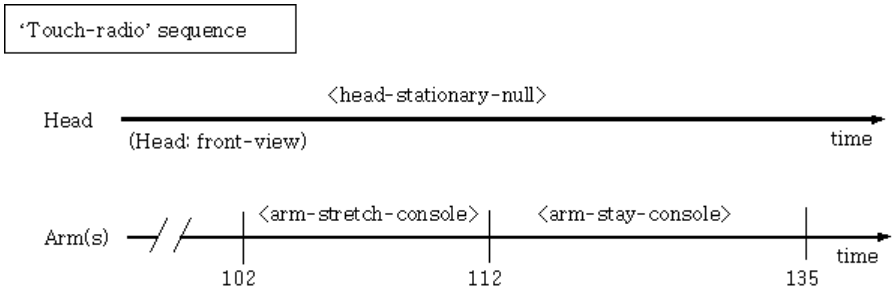


Figure 14: Semantic representation of *touch-radio* activity in Fig. 12 represented in terms of time-stamped operation triplets. The numbers on vertical axes denote the frame numbers. Note that the head orientation is in driver-centered viewpoint.

Figure 15: Segmented and tracked head in the sequence of color images

Acknowledgments

This work was partially supported by Volkswagen of America, Electronics Research Laboratory and by a grant from UC Discovery, Digital Media Innovations. We would also like to give a special thanks to our colleagues Dr. Tarak Gandhi and others in the Computer Vision and Robotics Research Laboratory for their invaluable inputs and assistance.

References

[1] S. Y. Cheng and M. M. Trivedi. Occupant posture modeling using voxel data:issues and framework. In *IEEE Proceedings on Intelligent Vehicles Symposium*, 2004.

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

[3] D. Gavrila and L. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Computer Vision and Pattern Recognition, 1996. Proceedings*, pages 73–80, 1996.

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

[5] K. Huang, M. M. Trivedi, and T. Gandhi. Driver's view and vehicle surround estimation using omnidirectional video stream. pages 444–449, 2003.

[6] D. Harwood K. Kim, T. H. Chalidabhongse and L. Davis. Background modeling and subtraction by codebook construction. In *IEEE International Conference on Image Processing (ICIP)*, 2004.

[7] J. McCall, O. Achler, and M. M. Trivedi. Design of an instrumented vehicle testbed for developing human centered driver support system. 2004.

[8] I. Mikic and M. M. Trivedi. Vehicle occupant posture analysis using voxel data. In *Ninth World Congress on Intellgient Trasport Systems*, October 2002.

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

[10] 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.

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

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

[13] M. M. Trivedi, S. Y. Cheng, E. C. Childers, and S. J. Krotosky. Occupant posture analysis with stereo and thermal infrared video. *IEEE Transactions on Vehicular Technology*, 2004.

[14] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.