Driver's View and Vehicle Surround Estimation using Omnidirectional Video Stream

Kohsia S. Huang, Mohan M. Trivedi, and Tarak Gandhi

Computer Vision and Robotics Research (CVRR) Laboratory
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, California, USA
khuang@ucsd.edu; mtrivedi@ucsd.edu; tgandhi@cvrr.ucsd.edu
http://cvrr.ucsd.edu/

Abstract

Our research is focused on the development of novel machine vision based telematic systems, which provide non-intrusive probing of the state of the driver and driving conditions. In this paper we present a system which allows simultaneous capture of the driver's head pose, driving view, and surroundings of the vehicle. The integrated machine vision system utilizes a video stream of full 360 degree panoramic field of view. The processing modules include perspective transformation, feature extraction, head detection, head pose estimation, driving view synthesis, and motion segmentation. The paper presents a multi-state statistical decision models with Kalman filtering based tracking for head pose detection and face orientation estimation. The basic feasibility and robustness of the approach is demonstrated with a series of systematic experimental studies.

1. Introduction

Driver distraction is an important issue in developing new generation of telematic systems [1]. To help reducing distractions caused by cell phone usage, a mobile machine vision system can be developed to actively control the conversation according to the driver status and the traffic conditions [2]. Our research is directed towards the development of a novel driver assistance system, "Visual Context Capture, Analysis and Televiewing (VCAT)." It derives visual context information on the driver and the traffic conditions. These cues could be used by the remote caller to change the conversational style according to events in or around the car, as shown in Figure 1. Visual cues about the driver and traffic conditions can be conveyed to the remote caller in raw video, in avatar and animated scene, and in cartoon formats. Thus the system provides a telepresence experience to the remote caller like a passenger sitting in the car. It also estimates the attentive direction of the driver and mitigates the conversation by audio-visual warnings. In this twofold effect, cell phone usage would be safer by avoiding the driver from being distracted.

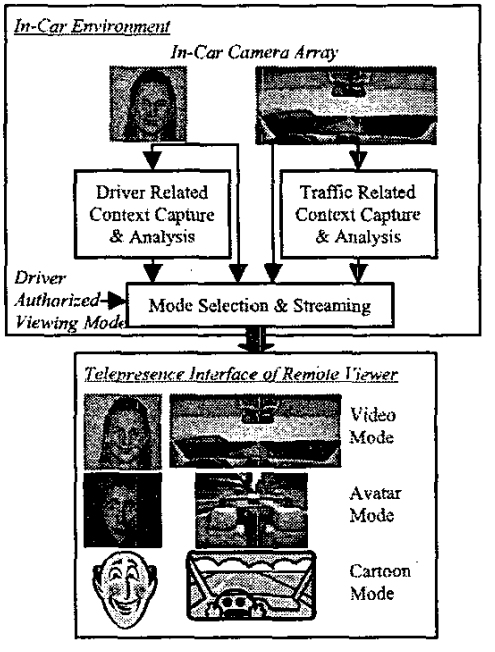

Figure 1. Information flow and context analysis of the VCAT system for driver assistance on cell phone safety.

In order to implement the VCAT system, a full coverage of the interior space and the dynamic scene outside of a vehicle must be captured for both televiewing and visual context analysis purposes. We use one omnidirectional camera, or omnicam, as the master sensor. The advantage of using omnicam is that it automatically supports event synchronization among in-vehicle and surroundings since they are captured in one shot. It can be used to extract preliminary visual context at lower resolution and higher processing speed, and possibly drive a few rectilinear cameras where higher resolution video is needed. As shown in Figure 2, multiple perspective views

This research is sponsored by UC Discovery (DiMI) Grants and Daimler Chrysler Corporation. The authors also want to thank their colleagues in the CVRR Laboratory for valuable assistance and cooperation.

can be simultaneously generated from the omnicam video on the driver, passengers, and surroundings by a nonlinear transformation with any pan, tilt, and zoom values [3]. This enables the VCAT system to analyze driver's viewing direction from the driver video and generate simultaneous driver's view from the omnicam video. The surroundings of the vehicle, including blind spots, can also be processed to estimate the traffic condition and detect potential risks. Using these information, the attentive status and workload of driver can be estimated, possibly with other information such as facial expression [4] and maneuvering of car [5]. This allows the VCAT system to decide when and how to mitigate cellular phone conversation and warn the driver appropriately.

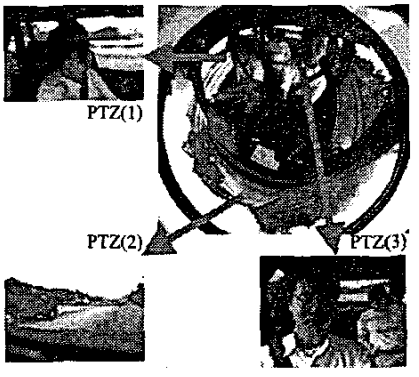

Figure 2. Simultaneous multiple perspective video generation on single omnidirectional video for event analysis. It enables frame-to-frame synchronization by nature.

In this paper we will cover two visual context analysis clusters for the driver's view generation and surround traffic conditions. Experimental evaluations on these modules will be presented.

2. Generation of Driver's View

In order to generate instantaneous driver's view, it needs to detect and track driver's head to extract driver's face, then estimate the driver's viewing direction from the driver's face image. Then the perspective view seen by the driver can be generated from the omni-video parallel to the driver's viewing direction.

2.1 Head Detection and Tracking

Head detection and tracking is a crucial module for the robustness of the driver assistant system. As compared to indoor situations [6], it is noted that (1) there is only one driver and the driver cannot wander around in car, and (2) the illumination condition is highly irregular both in intensity and in spectrum. For (1), we only need to generate a perspective view on the driver seat to find the driver's face. For (2), although skin-tone based face detection is the fastest, it will not be feasible here due to variant illumination spectrum. Edge-based methods are more robust then other feature extractions because they only rely on contrasts in the image. From the edge map, driver's head can be located by ellipse detection. The proposed head and face detection scheme is shown in Figure 3. A perspective view on the driver's seat is first generated. For faster processing, the image is sub-sampled and converted to gray scale for edge detection. Randomized Hough transform (RHT) [7] is used to search ellipses in the edge image with center, size, and orientation constraints to match general human heads. Each head candidate image is extracted by rotating the driver perspective image so that the corresponding ellipse aligns with a upright head pose in order to compensate head tilting. Driver's face image is cropped by a square window fitting to the ellipse and the image is scaled to a 64×64 image to reject non-face candidates by distance from feature space (DFFS) method [8]. Then the ellipse center, size, and orientation are used to update a set of constant velocity Kalman filters [9],

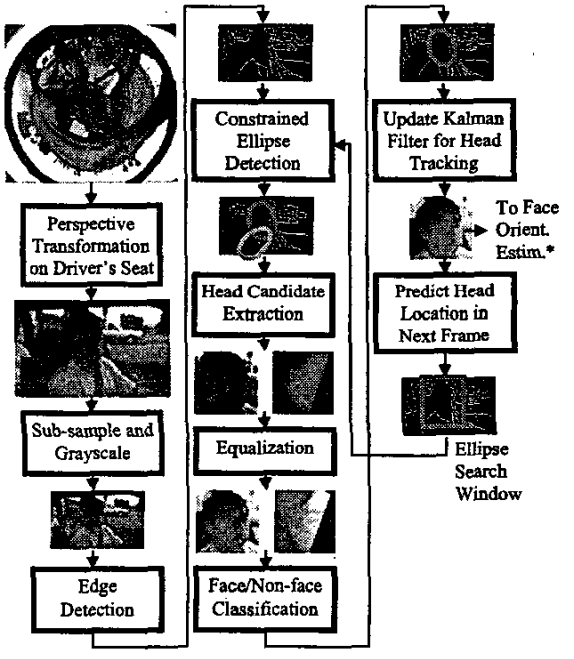

Figure 3. Driver's head detection and tracking.

$$\begin{bmatrix}\mathbf{x}(k+1)\\ \dot{\mathbf{x}}(k+1)\end{bmatrix}=\begin{bmatrix}\mathbf{I} & T\cdot\mathbf{I}\\ 0 & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}(k)\\ \dot{\mathbf{x}}(k)\end{bmatrix}+\begin{bmatrix}T^2\cdot\mathbf{I}/2\\ T\cdot\mathbf{I}\end{bmatrix}\nu(k)$$
$$\mathbf{y}(k)=\begin{bmatrix}\mathbf{I} & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}(k)\\ \dot{\mathbf{x}}(k)\end{bmatrix}+\omega(k) \quad (1)$$

where for ellipse center and size, state $\mathbf{x}$ and measurement $\mathbf{y}$ are 2 by 1 and $\mathbf{I}$ is 2 by 2 identity matrix. For ellipse orientation, $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{I}$ are 1 by 1. T is sampling interval or frame duration, i.e., 1/30 second. The covariance of measurement noise $\omega(k)$ is estimated from real-world data, and the covariance of random maneuver $\nu(k)$ is empirically

chosen by compromising between response time and sensitivity to noise. The states are used to interpolate detection gaps and predict the head position in the next frame. An ellipse search window is derived from the prediction and fed back to ellipse detection for the next frame. This window helps minimizing the area of ellipse search and reducing the epochs of RHT to increase the accuracy and speed. It also helps filtering false-positive head ellipses as in Figure 3.

The head tracking is initialized when an ellipse is detected and justified to be a head for some consecutive frames. Extensive RHT ellipse searching on the driver seat perspective view is used to find the first positive occurrence of head. Once driver's head is located and under tracking, the searching window is narrowed down and RHT uses less epochs to speed up the detection process. The track is terminated when no ellipse is detected and the predicted head location is classified as non-face for some consecutive frames.

2.2 Face Orientation Estimation and Driver's View Generation

The next step is to estimate driver's face orientation. The proposed method for the face orientation estimation is illustrated in Figure 4. After being adjusted for head tilting previously, driver's face image is compared to the view-based PCA templates to estimate the face orientation. In the training stage, we first collect a set of equalized training faces from the omnicam of multiple people with multiple horizontal face orientations. The orientation in the training faces varies approximately from –60 to 60 degrees with 30 degree step size. Then PCA subspace is constructed from the correlation matrix of the training faces [8] and all the training faces are projected into this subspace. Mean and covariance of the training projections are estimated for each face orientation category and a Gaussian likelihood function is approximated for each category. In the estimation stage, the scaled and equalized face image in the face video is projected into the PCA subspace and generates likelihood values on these five Gaussian distributions. The face orientation is thus estimated by maximum likelihood (ML). The estimated face orientation is then filtered by another Kalman filter as in equation (1). Then driver's viewing direction is computed from the filtered face orientation as in equation (2) and illustrated in Figure 4,

$$\begin{aligned} &\textit{Viewing Direction} = \\ &\quad (\textit{Direction of Driver}) - 180^\circ + \\ &\quad (\textit{Face Orientation}) \times K - \\ &\quad (x_{ellipse} - x_{perspective\ center}) \times (\text{degree per pixel}) \end{aligned} \tag{2}$$

where the constant K approximates the ratio of gazing direction to facing direction for empirical driver gazing behavior. The last term in equation (2) is used to take the exact location of head in the driver image into account, where $x_{ellipse}$ is the center of ellipse in x direction and $x_{perspective\ center}$ is the center of driver image in x direction. Thus driver's view video can be generated from the omnicam video with a fixed zooming factor to approximate human field of view, as shown in Figure 5.

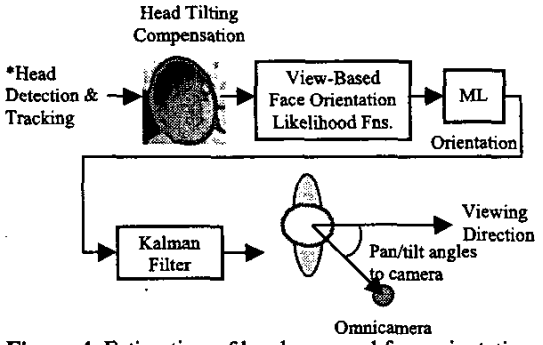

Figure 4. Estimation of head pose and face orientation, see text for details.

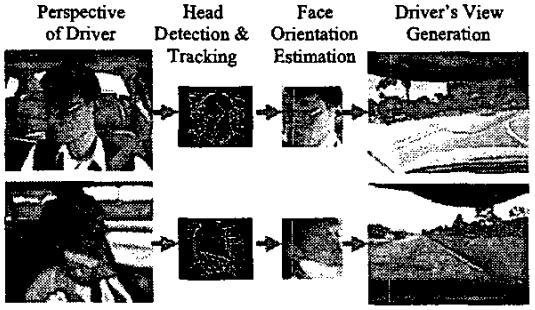

Figure 5. Some results of the perspectives of driver, constrained head detection and tracking, face orientation estimation, and instantaneous driver's view generation for televiewing. Note the differences in illumination condition and camera location in these video clips.

3. Experiment Results and Discussions of Driver's View Generation

Evaluation of the head tracking and face orientation estimation is accomplished using an extensive array of experimental data. We tested many driving video clips taken on different days and on different road, weather, and traffic conditions. Averaged head detection rates on two camera setups are summarized in Table 1. When low epoch RHT is applied without feedback of ellipse search window, the head detection rate is very poor. The rate improves if we use extensive RHT ellipse search on each frame, yet the processing speed is very slow. After the feedback loop is closed, we use extensive RHT search only on the first frame and fall back to rough RHT if the head is detected, the head detection rate is much improved to be as good as or even better than the extensive RHT, and the processing speed is as fast as the rough RHT. After KF tracking and interpolation, no frame is missed even in some tough situations like face occlusion, sharp uneven illumination, and turned-away face as shown in Figure 6.

The false positive rate is 9% if the DFFS bound is 2500 and is 7% if the DFFS bound is 2000.

Averaged Rate	Setup 1	Setup 2
Rough RHT, 1 Epoch	*32%*	*50%*
Rough RHT, 2 Epochs	*52%*	*61%*
Extensive RHT, 10 Epochs	*71%*	*79%*
RHT+Feedback, 10→1 Epochs	*64%*	*73%*
RHT+Feedback, 10→2 Epochs	*67%*	*87%*

Table 1. Averaged head detection rates before Kalman filtering on two camera setups. For setup 1, the camera is placed in front of the passenger seat and approximately 45° side viewing the driver. For setup 2, the camera is front-left to the driver. For rows 1 to 3, no ellipse search window is fed back and full image search is used. Note when search window is applied (row 4 and row 5), the detection rate of RHT ellipse search with less epochs is nearly as good as the rate of extensive RHT and the processing speed is much faster. After Kalman filtering, the head is latched on by the detected ellipse for all frames. DFFS bound for rejecting non-face candidates in these experiments is 2500.

Figure 6. Some situations that trouble the face orientation estimation.

Comparing setup 1 and setup 2 in Table 1, it suggests that the camera placement should be closer to the front of the driver. In this case the driver's face is more clear and the edge map of driver's head is closer to ellipse. Active infrared illumination would be helpful to increase head detection rate since it makes the image more clear and smoothes illuminations, weather, tunnel, and night variations. Also, there is a trade-off between head detection rate and speed for RHT based ellipse detection. Higher head detection rate would be desirable because the dynamics of head motion can be quickly captured. However, it would need more epochs and sacrifice the speed. It poses a need for less complicated ellipse detection algorithms. To further speedup the process, multiple processors or DSP hardware would be needed. The tasks of head detection and tracking in Figure 3 can be partitioned to view generation, edge detection, ellipse detection, and PCA-based face classification. Each part or a group of modules can be assigned to a specific processor.

Table 2 shows the averaged accuracies of face orientation estimation on test clips of different length. The error of face orientation estimation on each frame is compared to the manually estimated approximate ground truth value. The long, mid, and short term clips exhibit comparable accuracies. However for some situations as in Figure 6 and fast turning faces, the standard deviation of the estimation is about 3 to 4 times larger. For face occlusion, there is no good remedies except by interpolation along the frames using Kalman filter. The turned-away face could be alleviated by placing the omnicam near the front of the driver so it captures all the possible orientations of the face. For uneven illumination, PCA templates are prone to produce higher error rates. Other subspace feature analysis like LDA or ICA templates [10][11] would be helpful in this case.

Duration		Long	Mid	Short
Frames		*200*	*70*	*15*
Error Pre-KF	μ	*1°*	*0°*	*-1°*
	σ	*7°*	*19°*	*8°*
Error Post-KF	μ	*0°*	*4°*	*-1°*
	σ	*8°*	*7°*	*7°*

Table 2. Long, mid, and short term accuracies of face orientation estimation. The face video is cropped by a closed-loop head detection and tracking with RHT of 10→2 epochs. The error before KF is the error of the output of the ML face orientation estimation and the error after KF is the error after the Kalman filtering in Figure 4.

Eye-gaze direction estimation is needed for an accurate driving view. In equation (2), we use a rough estimate of driver's gazing direction from driver's face orientation. Rectilinear camera set on the dash board would be needed because the omnicam resolution is not sufficient for the pupil. A commercial system, faceLab, of Seeing Machines is an example for this purpose [12]. Also, active infrared illumination could be useful to estimate eye-gaze direction by bright pupil effect.

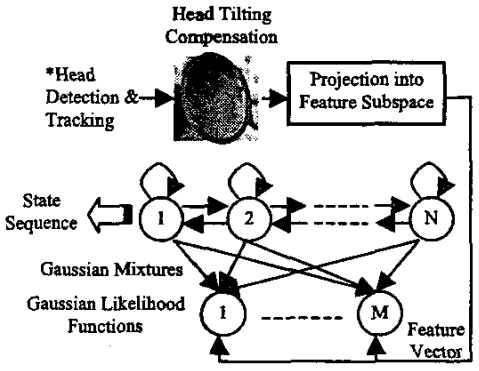

Figure 7. Modified face orientation estimation by continuous density HMM. Face video is projected into feature subspace and generates *M* Gaussian likelihood values.

To improve the dynamic performance of face orientation estimation, we observe that the Gaussian likelihood information in Figure 4 is discarded by the maximum likelihood decision. We can construct a

continuous density HMM to fully utilize these information, as shown in Figure 7 [2]. The Markov chain could have N=13 states which represent face orientations from approximately –90 to 90 degrees with 15 degree step size. The observation probability of the j-th states $b_j(O)$ can be modeled by a mixture of the five Gaussian distributions [13] in PCA subspace for each training face orientation category as previously mentioned. The state sequence $q(k)$ given a driver's face video can be estimated by maximum *a posteriori* (MAP) estimation in real-time or optimally estimated by Viterbi algorithm [13] with some delay caused by sequence framming. The estimated state sequence represents the face orientation movement of the driver. Although the experimental data are still under evaluation, we anticipate that this approach will out-perform the method in Figure 4 in that it is a *delayed decision* approach [14]. The face orientation sequence can be further utilized to estimate driver's attentive and psychological status by a hierarchical layer of estimators such as Bayesian nets [15]. We will conduct experiments on these schemes in the near future.

4. Surround Monitoring

In addition to monitoring the driver's state, visual context of surround traffic conditions can also be derived from the omni-video. In order to detect the nearby vehicles, simple background subtraction for indoor environments would not be feasible since the car is always moving. In this section we present a motion-based surround vehicle segmentation scheme suitable for the mobile VCAT application.

From the camera on the vehicle, the independent motion of the surrounding vehicle would be separated from the ego-motion of the road. The road can be modeled as a planar surface, and the motion of the points on the road is given by the following transformation [16],

$$(x' \quad y' \quad z')^T = H\,(x \quad y \quad z)^T \tag{3}$$

where $(x \quad y \quad z)$ and $(x' \quad y' \quad z')$ are the homogenous image coordinates of a point on the road in two frames, and H is a 3×3 matrix expressed in terms of camera motion and calibration. Features that are not on the road or have independent motion do not satisfy this model, and have residual motions. Thus a surrounding vehicle can be detected by warping one image to another and comparing the motion compensated images.

For omnidirectional video, the image distortion due to camera optics also needs to be considered. A combined transformation for mapping a point in one image to another is given by the following procedure:

(1) Convert the pixel coordinate from the omni-image to the homogenous coordinates in the perspective view by $(x \quad y \quad z)^T = f\left((x_p \quad y_p)^T\right)$, (cf. [3])

(2) Apply the transformation (3) to compensate ego-motion, and

(3) Convert the motion compensated point back to the omni-image by $(x'_p \quad y'_p)^T = f^{-1}\left((x' \quad y' \quad z')^T\right)$.

Approximate estimate of the planar motion transformation H is obtained from the camera calibration parameters, as well as from the vehicle speed using the CAN bus. However, if the camera is vibrating, or its velocity is inaccurately known, features on the road surface such as lane marks could also have residual motion. This residual motion can be used to refine the motion parameter estimates in a Bayesian framework [16]. Here, the approach is generalized for omnicameras to optimally combine the prior knowledge of motion parameters with the motion residual of the image features.

Under favorable conditions, the spatial gradients (g_x, g_y), the temporal gradient (g_t), and the image velocity $(u_x, u_y) = (x'_p - x_p, y'_p - y_p)$ of an image satisfy the optical flow constraint,

$$g_x u_x + g_y u_y + g_t = 0 \tag{4}$$

Image motion is expressed parametrically in terms of motion parameters for a number of image points as $\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{v}$ where

$$\begin{aligned} \mathbf{x} &= (H_{11} \quad H_{12} \quad \ldots \quad H_{32})/H_{33} \\ \mathbf{h}(\mathbf{x}) &= g_x(x'_p - x_p) + g_y(y'_p - y_p), \quad \mathbf{z} = g_t \end{aligned} \tag{5}$$

and $\mathbf{v}$ is the measurement noise in the time gradient. The estimates of the state $\mathbf{x}$ and its covariance $\mathbf{P}$ are iteratively updated using the measurement update equations of the iterated extended Kalman filter [9],

$$\begin{aligned} \mathbf{P}_{i+1} &= \left(\mathbf{H}_i^T \mathbf{R}^{-1} \mathbf{H}_i + \mathbf{P}_-^{-1}\right)^{-1} \\ \hat{\mathbf{x}}_{i+1} &= \hat{\mathbf{x}}_i + \mathbf{P}_{i+1}\left[\mathbf{H}_i^T \mathbf{R}^{-1}\left(\mathbf{z}_i - \mathbf{h}(\mathbf{x}_i)\right) - \mathbf{P}_-^{-1}(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_-)\right] \end{aligned} \tag{6}$$

where $\mathbf{H}_i$ is the Jacobian of $\mathbf{h}(\mathbf{x})$ at $\mathbf{x} = \hat{\mathbf{x}}_i$, $\mathbf{x}_-$ is the prior approximately known state, and $\mathbf{P}_-$ is the prior covariance. To avoid using outlier points that do not satisfy the ego-motion model, robust estimation [17] is applied. Also, since the optical flow constraint is valid only for small image motion, coarse to fine estimation [18] is used.

Figure 8 shows the experimental results of the surrounding vehicle detection. The image motion is analyzed in the area of interest on the driver side. Normalized difference [19] between the motion compensated images is used to enhance the vehicle features having height or independent motion and to attenuate the road features. Post-processing is used to further suppress remaining road features, and the components that are close to each other are grouped into a bounding box.

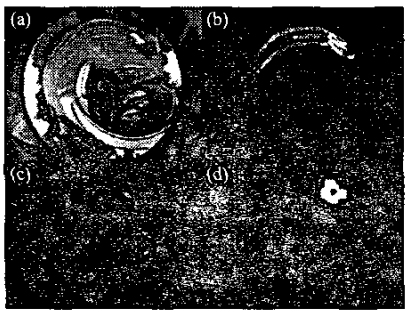

Figure 8. Result of surrounding vehicle detection. (a) Current frame of the image, with estimated image motion in the area of interest. (b) Points used for estimation of ego-motion. Gray: inliers, White: outliers, Black: unused. (c) Normalized frame difference in the area of interest. (d) Output after post-processing and clustering.

The robustness of the proposed scheme can be further improved by integrating the estimation process over frames. For example, outlier removal could be improved by propagating the outlier pixels from frame to frame. Similarly, the motion parameters can be updated over time using Kalman filter. For better driver assistance, other detection modules for lanes, pedestrians, and traffic signs could also be added.

5. Concluding Remarks

In this paper we have presented the VCAT driver assistance system in order to enhance cell phone safety for the driver. We described development of an integrated machine vision system for accurate and robust estimation of the driver's viewing direction and surrounding traffic conditions using an omnicamera. Novel algorithms using Kalman filtering based tracker and multi-state HMM models have been evaluated using a series of experimental studies. These experiments have proven the basic feasibility and promise of the approach. Enhancement of the system performance can be accomplished by using higher resolution video, specialized in-vehicle illumination, and embedded processors. In the future, these modules could be integrated for the estimation of driver's attentive status and workload.

References

[1] Rosalyn G. Millman, *Keynote Address, National Highway Traffic Safety Administration (NHTSA) Public Meeting on Driver Distraction*, Jul. 18, 2000, Washington, D.C.

[2] K. S. Huang and M. M. Trivedi, "Driver Head Pose and View Estimation with Single Omnidirectional Video Stream," *Proc. 1st. Int'l. Workshop on In-Vehicle Cognitive Computer Vision Systems*, Graz, Austria, Apr. 3, 2003.

[3] K. S. Huang and M. M. Trivedi, "Video Arrays for Real-Time Tracking of Persons, Head and Face in an Intelligent Room," To appear in *Machine Vision and Applications, Special Issue*, Jun. 2003.

[4] J. C. McCall, S. P. Mallick, and M. M. Trivedi, "Real-time Driver Affect Analysis and Tele-Viewing System," *Proc. IEEE Intelligent Vehicle Symposium*, Columbus, OH, USA, Jun. 9-11, 2003.

[5] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," *Proc. IEEE Conf. CVPR.*, pp. 994-999, 1997.

[6] M. Trivedi, K. Huang, and I. Mikic, "Intelligent Environments and Active Camera Networks," *Proc. IEEE Systems, Man and Cybernetics Conf.*, vol. 2, pp. 804-809, Oct. 2000.

[7] R. McLaughlin, "Randomized Hough Transform: Better Ellipse Detection," *Proc. IEEE TENCON Digital Signal Processing Applications*, pp. 409-414, 1996.

[8] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Comp. Soc. Conf. on Comp. Vis. and Patt. Recog.*, pp. 586-591, Jun. 1991.

[9] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation*, John Wiley and Sons, 2001.

[10] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE. Trans. PAMI*, vol. 19, no. 7, pp. 711-720, Jul. 1997.

[11] G. L. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying Facial Actions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-989, 1999.

[12] faceLab, http://www.seeingmachines.com/.

[13] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[14] K. Huang and M. Trivedi, "Streaming Face Recognition using Multicamera Video Arrays," *Proc. Int'l Conf. on Patt. Recog.*, vol. 4, pp. 213-216. Aug. 2002.

[15] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer-Verlag, 2001.

[16] W. Kruger, "Robust real time ground plane motion compensation from a moving vehicle," *MVA*, vol. 11, pp. 203-212, Nov. 1999.

[17] G. Danuser and M. Stricker, "Parametric model fitting: From inlier characterization to outlier detection," *IEEE Trans. PAMI*, vol. 20, no. 2, pp. 263-280, Mar. 1998.

[18] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields" CVIU, vol. 63, no. 1, pp. 75-104, 1996.

[19] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*, Prentice Hall, 1st Ed., Mar. 1998.