Appeared in Proceedings of the 1st International Workshop on In-Vehicle Cognitive Computer Vision Systems,
in conjunction with the 3rd International Conference on Computer Vision Systems,
Graz, Austria, pp. 44-51, April 3, 2003.

DRIVER HEAD POSE AND VIEW ESTIMATION WITH SINGLE OMNIDIRECTIONAL VIDEO STREAM

Kohsia S. Huang and Mohan M. Trivedi

Computer Vision and Robotics Research (CVRR) Laboratory
University of California, San Diego
http://cvrr.ucsd.edu/

ABSTRACT

Driver distraction is an important issue in developing new generation of telematic systems. Our research is focused on development of novel machine vision systems, which can provide better understanding of the state of the driver and driving conditions. In this paper we discuss in detail on the development of a system which allows simultaneous capture of the driver's head pose and driving view. The system utilizes a full 360 degree panoramic field of view using a single video stream. The integrated machine vision system includes modules of perspective transformation, feature extraction, head detection, head pose estimation, and driving view synthesis. 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 the help of a series of systematic experimental studies.

Keywords: driver assistance system, closed-loop head detection and tracking, face orientation estimation, driver's view generation.

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 talking 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 load 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.

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 video 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 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 also generate simultaneous driver's view from the omnicam video. Using these videos, 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 to mitigate cellular phone conversation. Meanwhile, with the analysis of the surrounding traffic conditions, the system can detect potential risks at which the driver is not paying attention and warn the driver appropriately.

In this paper we focus on the visual context analysis module which generates instantaneous driver's view.

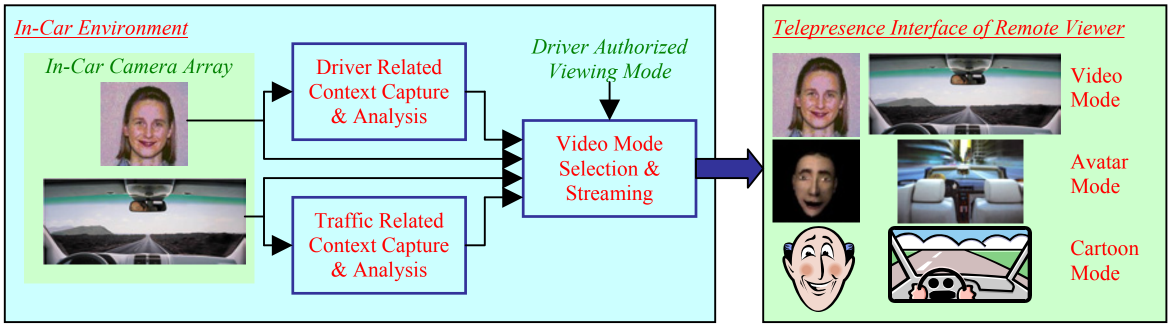
In-Car Environment
In-Car Camera Array
Driver Related Context Capture & Analysis
Traffic Related Context Capture & Analysis
Driver Authorized Viewing Mode
Video Mode Selection & Streaming
Telepresence Interface of Remote Viewer
Video Mode
Avatar Mode
Cartoon Mode

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

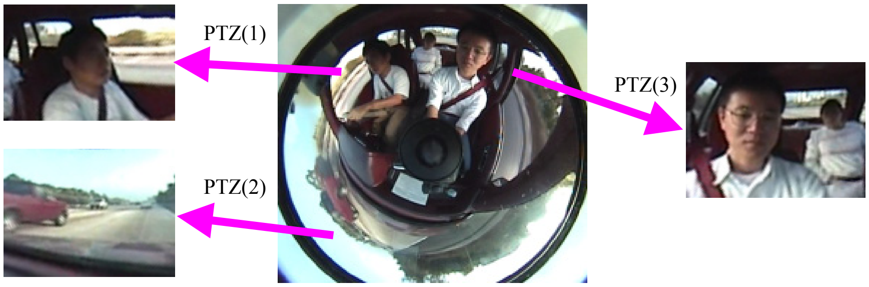

Figure 2. Simultaneous multiple perspective video generation on single omnidirectional video for event analysis. It enables frame-to-frame synchronization by nature.

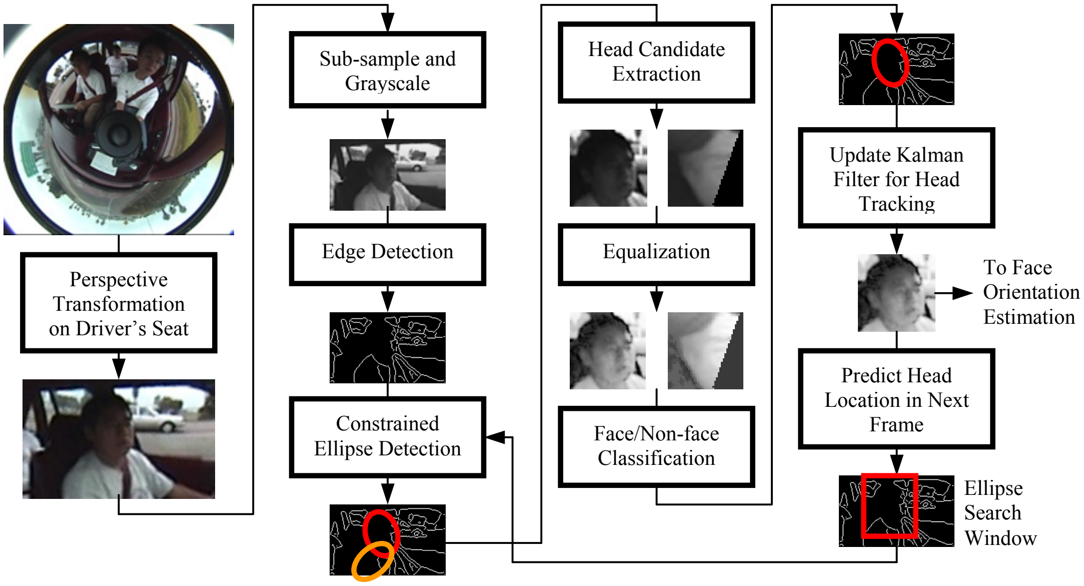
Perspective Transformation on Driver's Seat
Sub-sample and Grayscale
Edge Detection
Constrained Ellipse Detection
Head Candidate Extraction
Equalization
Face/Non-face Classification
Update Kalman Filter for Head Tracking
To Face Orientation Estimation
Predict Head Location in Next Frame
Ellipse Search Window

Figure 3. Driver's head detection and tracking.

Experimental evaluations and discussions on this module are then 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 driver's face image. Then the perspective view seen by driver can be generated from the omni-video parallel to driver's viewing direction.

2.1 Head Detection and Tracking

The design consideration of head detection and tracking for in-vehicle application is on robustness and speed, and they are somewhat contradict to each other. 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 some center, size, and orientation limitations on the ellipses 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][9]. Then the ellipse center, size, and orientation are used to update a set of constant velocity Kalman filters [10],

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

where for ellipse center and size, state **x** and measurement **y** are 2 by 1 and **I** is 2 by 2 identity matrix. For ellipse orientation, **x**, **y**, and **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 searching and less extensive RHT can be used, therefore increases 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 head pose. The proposed method for head pose estimation is illustrated in Figure 4. Driver's face image from head detection and tracking has been adjusted for head tilting. The image is then compared to the view-based PCA templates to estimate the face orientation. We first collect a set of equalized training faces of multiple people with multiple horizontal face orientations [11] from the omnicam. 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 [12] and all the training faces are projected into this subspace. Mean and covariance of the projections are estimated for each face orientation category and a Gaussian distribution is approximated for each category. As compare to [12], face orientations are categorized instead of the identities of people. 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 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 and driver's direction to the omnicam as in equation (2) and illustrated in Figure 4,

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

where the facing direction is in terms of 0° of the omnicam and is the pan factor used to generate driver's perspective view from the omni video. 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

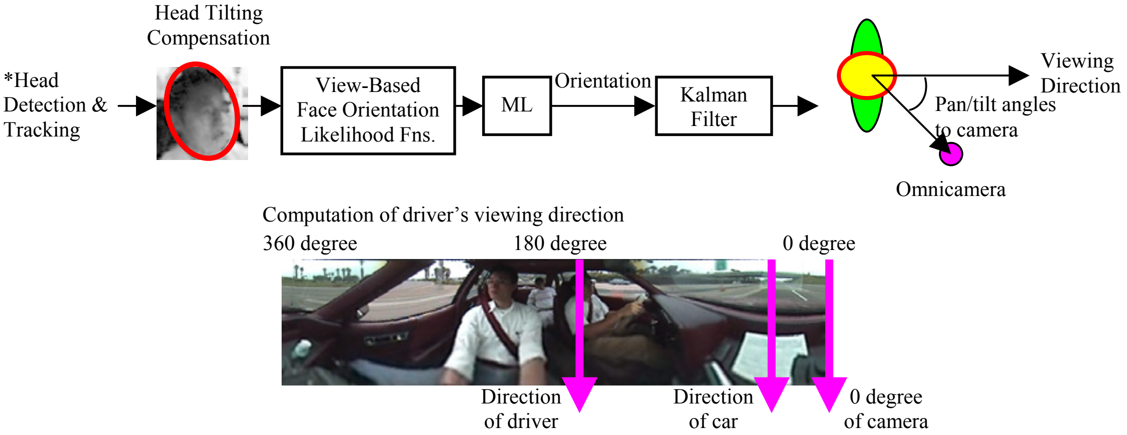

Figure 4. Estimation of head pose and face orientation, see text for details. The direction of driver to the camera is involved in the estimation of viewing direction.

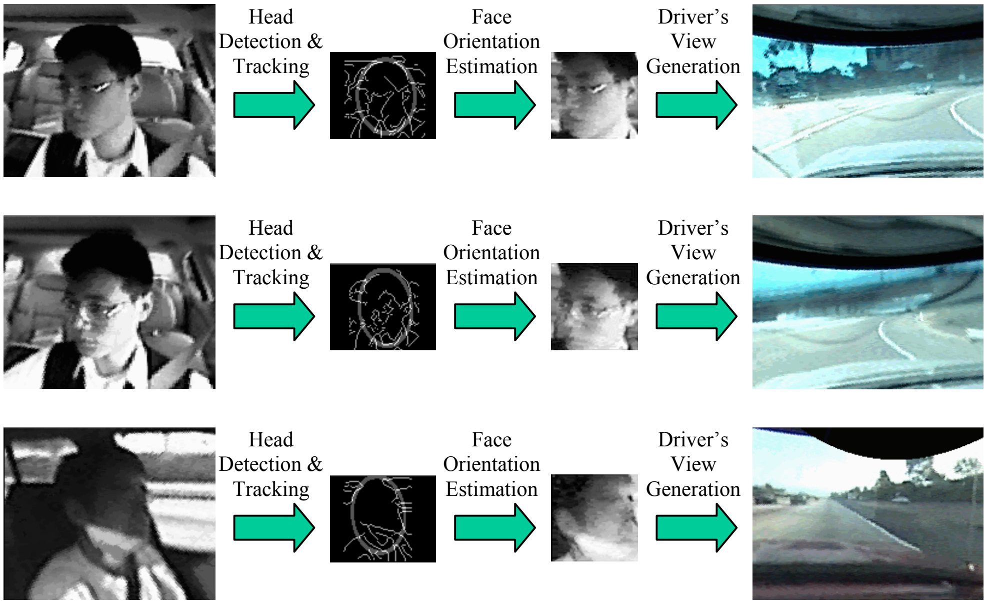

Figure 5. Perspective of driver, constrained head detection and tracking, face orientation, and instantaneous driver's view generation for televiewing. Note the differences in illumination condition and camera location in these video clips.

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.

3. EXPERIMENT RESULTS AND DISCUSSIONS

Evaluation of the performance of head tracking and face orientation estimation is accomplished using an extensive array of experimental data. One set of video clips is collected earlier with the omnicam set on a tripod sitting on the floor of the passenger seat. The clips are taken on different times in the day and on different road, weather, and traffic conditions. Head detection rates on the older and newer video clips are summarized in Table 1 and Table 2 respectively. RHT head detection rate is the ratio of frames where the head ellipse is detected to the total number of frames in the video clip. When rough RHT is applied without feedback of ellipse search window, head detection rate is pretty low. The rate improves if we use extensive RHT ellipse search on each frame. However the processing speed is very slow. After

Clip	Frames	Rough RHT, 1 Epoch	Rough RHT, 2 Epochs	Extensive RHT, 10 Epochs	RHT+ Feedback, 10→1 Epochs	RHT+ Feedback, 10→2 Epochs
#1	*200*	*33%*	*58%*	*69%*	*63%*	*67%*
#2	*75*	*29%*	*45%*	*75%*	*68%*	*67%*
Avg.	—	32%	52%	71%	64%	67%

Table 1. Head detection rates before Kalman filtering of 2 video clips. The camera is placed in front of the passenger seat and approximately 45° side viewing the driver. For columns 3 to 5, no ellipse search window is fed back and full image search is used. Note when search window is applied, 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 filter, the head is latched on by the detected ellipse for all frames. DFFS bound for rejecting non-face candidates in these experiments is 2500.

Clip	Frames	Rough RHT, 1 Epoch	Rough RHT, 2 Epochs	Extensive RHT, 10 Epochs	RHT+ Feedback, 10→1 Epochs	RHT+ Feedback, 10→2 Epochs
#3	*15*	*53%*	*67%*	*84%*	*80%*	*91%*
#4	*15*	*40%*	*42%*	*71%*	*62%*	*71%*
#5	*15*	*58%*	*76%*	*80%*	*76%*	*98%*
Avg.	—	50%	61%	79%	73%	87%

Table 2. Head detection rates before Kalman filtering of 3 driver video clips. The camera is placed in front-left of the driver. Note when search window is applied, the detection rate of RHT ellipse search with less epochs is even better than the rate of extensive RHT and the processing speed is much faster. After Kalman filter, the head is latched on by the detected ellipse for all frames. DFFS bound for rejecting non-face candidates in these experiments is 2500.

DFFS Bound	False Positive Rate
2500	*9%*
2000	*7%*

Table 3. False positive rate of head detection before Kalman filtering. The head detection uses closed-loop RHT ellipse search with 10→2 epochs. One video clip of empty driver seat is repeatedly tested under different values of DFFS bound.

the feedback loop is closed, we use extensive RHT search only on the first frame and fall back to rough RHT if 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 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. Table 3 shows the false positive rates under different DFFS settings.

Comparing Table 1 and Table 2, it suggests that the camera placement should be closer to 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 driver image more clear and smoothes uneven illuminations, weather, tunnel, and night situations. 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 reflected in head tracking and face orientation estimation. However, it would require more epochs and sacrifice real-time requirement. 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 4 and Table 5 shows the accuracies of face orientation estimation on different video clips of different length. The error of face orientation estimation on each frame is compared to the approximate ground truth value estimated by human. Both the short term and long term

Clip	Frames	Approximate Ground Truth	Error before KF		Error after KF		Note
			Mean	Std. Dev.	Mean	Std. Dev.	
#1	*200*	*35°~23°~35°*	*-1°*	*8°*	*-1°*	*7°*	
#2	*75*	*35°*	*-19°*	*27°*	*-18°*	*24°*	Sharp uneven illumination
#3	*70*	*35°*	*1°*	*7°*	*0°*	*8°*	
#4	*30*	*35°*	*16°*	*28°*	*-15°*	*16°*	Face occluded

Table 4. Mid-to-long term accuracy of face orientation estimation. The camera is placed in front of the passenger seat and approximately 45° side viewing the driver. The face video is cropped by a closed-loop head detection and tracking with RHT of 10→2 epochs. Error before KF is the error of the output of ML face orientation estimation and error after KF is the error after Kalman filter as in Figure 4.

Clip	Frames	Approximate Ground Truth	Error before KF		Error after KF		Note
			Mean	Std. Dev.	Mean	Std. Dev.	
#5	*15*	*-25°*	*0°*	*19°*	*4°*	*7°*	
#6	*15*	*-25°*	*-3°*	*8°*	*-2°*	*3°*	
#7	*15*	*0°~70°*	*-45°*	*32°*	*-50°*	*17°*	Rapid face turning

Table 5. Short term accuracy of face orientation estimation. The camera is placed in front-left of the driver. The face video is cropped by a closed-loop head detection and tracking with RHT of 10→2 epochs.

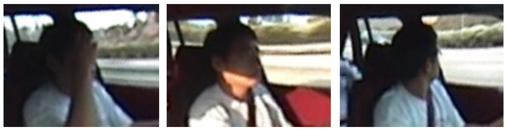

Figure 6. Some situations that trouble the face orientation estimation.

clips exhibit comparable accuracies. However for some problematic situations like in Figure 6, the face orientation estimation shows big error deviation. For face occlusion, there is no good way to estimate the orientation except by interpolation along the frames using Kalman filter. The turned-away face problem could be alleviated by placing the omni-camera near the front of the driver so it captures all possible orientations of the face. For uneven illumination situation, PCA templates are prone to produce higher error rates. Other subspace feature analysis like LDA or ICA templates [13][14][15] would be helpful in this case. We will discuss on a better scheme to improve the face orientation estimation.

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 [16]. Also, active infrared illumination on driver's face could be useful to estimate eye-gaze direction by bright pupil effect.

To improve the performance of face orientation, we propose another scenario similar to [12]. We can construct a continuous density HMM with 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 in PCA subspace for each training face orientation category as previously mentioned, or more generally M Gaussian mixtures,

$$b_j(O) = \sum_{m=1}^{M} c_{jm} N(O, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}) \tag{3}$$

where O is the projection vector of the adjusted face image in feature subspace, c_{jm}, $1 \le j \le N$ is the mixture coefficient which sums up 1 on m, and $\boldsymbol{\mu}_{jm}$ and $\mathbf{U}_{jm}$ are the mean and covariance of the Gaussian density, respectively. The HMM is illustrated in Figure 7. The state sequence $q(k)$ given a driver's face video can be estimated by maximum *a posteriori* (MAP) estimation in real-time as

$$q(k) = \arg\max_{1 \le j \le N} b_j\big(O(k)\big) P\big(q(k) = S_j \big| q(k-1)\big) \tag{4}$$

or optimally estimated by Viterbi algorithm [17] with some delay caused by sequence framming. The initial probability π and state transition probability A of the hidden Markov chain as well as the parameters in equation (3) are estimated by the EM algorithm [18]. Video clips of driver's face should be collected and projected into feature subspace to carry out the HMM distribution

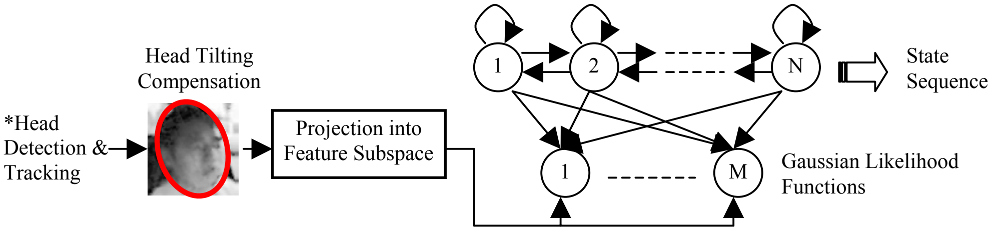

Figure 7. Modified face orientation estimation by continuous density HMM. Face video is projected into feature subspace and generates *M* Gaussian likelihood values. Those values are observed by each state and a state sequence can be estimated to describe the face orientation sequence in the video in some optimal sense.

parameter estimations. Currently the experimental results are not availble, yet we anticipate that this approach to face orientation estimation will out-perform the previous method in that it is a *delayed decision* approach. The weakness of the previous method is that before Kalman filtering, the useful likelihood information is discarded by maximum likelihood decision. The estimated state sequence represents the face orientation movement of the driver. Continuous state HMM such as Kalman filter with likelihood functions as observations is also of interest to develop for higher resolution description of the facing movement. The face orientation motions can be further utilized to estimate driver's attentive and psychological status by a hierarchical layer of estimators such as Bayesian nets [19]. We will conduct experiments on these schemes in the near future.

4. CONCLUDING REMARKS

In this paper we have proposed the VCAT driver assistance system in order to enhance cell phone safety for the driver. The in-vehicle system modules focus on the recognition of driver status, which include head detection and tracking for driver face extraction and estimation of instantaneous driver's view for assessing driver's attentive focus and cognitive load. In this paper we described development of an integrated machine vision system for accurate and robust estimation of the driver's view using a single omni video stream. 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.

ACKNOWLEDGEMENTS

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.

REFERENCES

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

[2] K. S. Huang, M. M. Trivedi, and T. Gandhi, "Driver's View and Vehicle Surround Estimation using Omnidirectional Video Stream," To appear in *IEEE Intelligent Vehicle Symposium*, Columbus, OH, USA, Jun. 9-11, 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," To appear in *IEEE Intelligent Vehicle Symposium*, Columbus, OH, USA, June 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. M. Trivedi, K. S. Huang, and I. Mikic, "Intelligent Environments and Active Camera Networks," *Conf. Proc. IEEE Systems, Man and Cybernetics*, 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] E. Hjelmas and B. K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding* **83**, pp. 236-274, 2001.

[9] 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.

[10] Y. Bas-Shalom and T. Fortmann, *Tracking and Data Association*, Acadamic Press, 1988.

[11] D. Beymer, "Face Recognition Under Varying Pose," *Proc. 1994 IEEE Comp. Soc. Conf. on Comp. Vis. and Patt. Recog.*, pp. 756-761, Jun. 1994.

[12] K. S. Huang and M. M. Trivedi, “Streaming Face Recognition using Multicamera Video Arrays,” *Proc. Int’l Conf. on Patt. Recog.*, Vol. 4, pp. 213-216, Aug. 2002.

[13] K. Etemad and R. Chellappa, “Discriminant Analysis for Recognition of Human Face Images,” *J. Opt. Soc. Am. A*, Vol. 14, No. 8, pp. 1724-1733, Aug. 1997.

[14] 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.

[15] 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.

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

[17] 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.

[18] A. Dempster, N. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *J. Roy. Stat. Soc. B*, Vol. 39, No. 1, pp. 1-38, 1977.

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