2012 15th International IEEE Conference on Intelligent Transportation Systems
Anchorage, Alaska, USA, September 16-19, 2012

Optical flow based Head Movement and Gesture Analysis in Automotive Environment

Sujitha Martin, Cuong Tran, Ashish Tawari, Jade Kwan and Mohan Trivedi

Abstract— **Head gesture detection and analysis is a vital part of looking inside a vehicle when designing intelligent driver assistance systems. In this paper, we present a simpler and constrained version of Optical flow based Head Movement and Gesture Analyzer (OHMeGA) and evaluate on a dataset relevant to the automotive environment. OHMeGA is user-independent, robust to occlusions from eyewear or large spatial head turns and lighting conditions, simple to implement and set-up, real-time and accurate. The intuitiveness behind OHMeGA is that it segments head gestures into head motion states and no-head motion states. This segmentation allows higher level semantic information such as fixation time and rate of head motion to be readily obtained. Performance evaluation of this approach is conducted under two settings: controlled in laboratory experiment and uncontrolled on-road experiment. Results show an average of 97.4% accuracy in motion states for in laboratory experiment and an average of 86% accuracy overall in on-road experiment.**

I. INTRODUCTION

The World Health Organization (WHO) predicted that road traffic injuries would become the 3rd leading cause of global burden of diseases as rank ordered by DALYs (disability-adjusted life year) by 2020; not to mention the economic cost of global road crashes was estimated at US $518 billion [11]. The emotional and economic burdens on households due to these tragedies heighten the need for safer automobiles. One way to make automobiles safer is to incorporate Intelligent Driver Assistance Systems (IDAS) into our vehicles to warn against or to help mitigate dangerous situations.

Dangerous situations need to be assessed not only by looking outside the vehicle but also by looking inside the vehicle [15] because if the driver is already aware of impending danger outside the vehicle, IDAS should focus on how to mitigate rather than warn against the impending danger. There are multiple objects of interest when looking inside the vehicle, especially those concerned with the driver operating the vehicle, such as the driver's eye gaze, head gesture, upper body posture, hand positions, and foot movements. Fig. 1 shows scenarios where knowing head positions and head movements give vital information on the driver's focus of attention and intent to maneuver. In this paper, we focus on head gestures and show an intuitive way of inferring and deriving inherent properties of head gestures.

First, we provide a systematic interpretation of head gesture in terms of head pose and head dynamics. Head pose is well defined as the 3D-orientation of a head with it's 3 degrees of freedom relative to a fixed world coordinate system. Given head pose at time t and $(t + \triangle t)$, head dynamics is the motion that describes the change in head position in the time $\triangle t$ duration. Whereas head dynamics encodes where the head moved relative to a starting position, head gesture encodes how the head moved from the starting position to the ending position. Head gesture then is denoted by a combination of head motions and head positions over a period of time.

Authors with the Laboratory for Intelligent and Safe Automobiles(cvrr.ucsd.edu/LISA), University of California, San Diego, La Jolla, CA 92037 {scmartin, cutran, atawari, jade, mtrivedi}@ucsd.edu

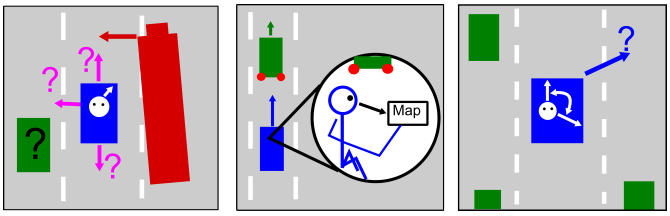

Fig. 1. Example scenarios where knowledge of the driver's head orientation and motion can help to infer possible dangerous situations: (a) driver is focused on the truck wavering into his lane but may not be aware of the vehicle in his left lane, (b) driver is focused on his navigation system and does not notice the car braking in front of him, (c) driver wishes to change into the right lane and turns his head towards the prospective lane to check for vehicles occupying the prospective lane.

Ideally a history of continuous head pose estimation will be sufficient to describe a head gesture. In real-world driving, however, estimating head pose in a continuous manner is a challenging task, not to mention computationally intensive [10] [17]. Due to limitations such as manual calibration of camera sensors for each user and loss of head tracking from occlusions of facial features due to eyewear or large spatial head turns, head gesture analysis using head pose alone is not desired in driving scenarios.

We present a simpler and constrained version of an approach called OHMeGA [8], which is user-independent and robust to occlusions from eyewear or large spatial head turns and lighting conditions, yet utilizes a simple set-up (e.g. a frontal facing monocular camera). Furthermore, OHMeGA runs in real-time, an important requirement of IDAS. The intuitiveness of OHMeGA is that head gestures can be broken down into moving states (i.e. head motions) and fixation states (i.e. no head motions) as shown in Fig. 2. This simpler version of OHMeGA takes advantage of the fact that drivers are fixated straight most of the time and the fact that when using frontal facing cameras, head motions in the pitch and yaw rotation angles translate to motions in the x-and y-directions of the image plane. By segmenting head gestures into move states and fixation states, higher level semantic information such as fixation time and rate of head motion

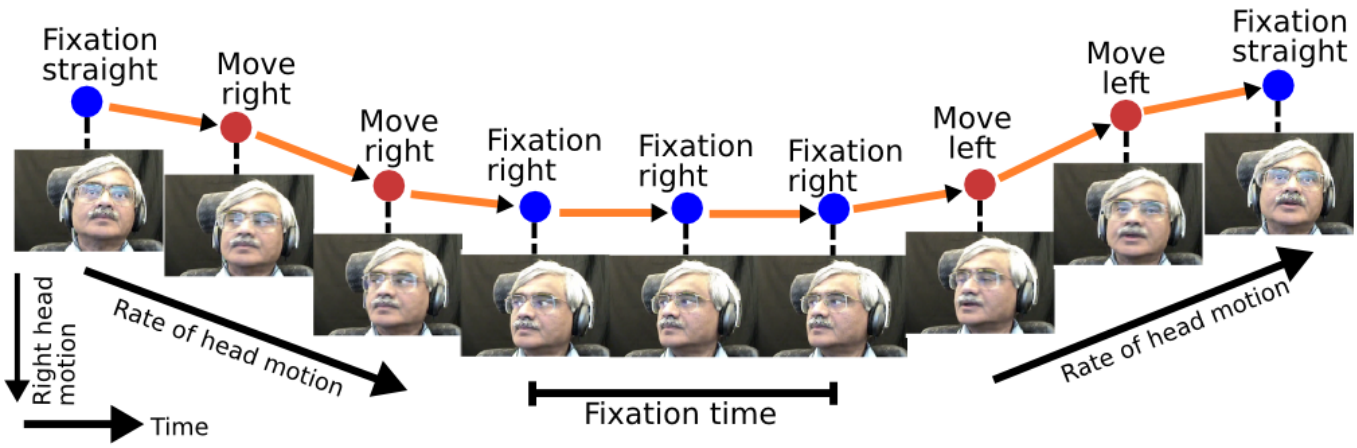

Fig. 2. Illustration of typical head movements, gestures, fixations and temporal dynamics.

can be readily obtained.

The remainder of this paper is organized as follows. Section II provides studies related to head gesture analysis for driver safety and other applications. Detailed description on the OHMeGA approach and how to overcome implementation difficulties are discussed in Section III. Section IV describes the experiments and provides quantitative evaluation of our approach. Finally, Section V concludes the discussion with future directions.

II. Related Studies

Many research groups have contributed significant work in the field of gesture recognition . In automotive environment, however, study of head gesture has gained increasing attention in recent years. In [9], Morris et al. uses higher level semantic knowledge on head gestures as feature vectors for lane change intent prediction. Even though Cheng et al. [2] doesn't use "expert" knowledge of the head, the author uses information on the driver's head pose for turn intent analysis. For determining vigilance [1] and driver fatigue [19], Ji et al. and Bergasa et al studied head nodding frequency using head pose. Head pose and head movements have also been useful for active displays using holistic sensing [16] and for attention estimation [4].

In other application domains, many studies have been conducted on head gesture recognition and analysis. Among these applications, studies of fixation on a person, a scene or an object has been of particular interest in surveillance [12] and meeting like scenarios [13], where it's possible to also infer joint attention of a group of individuals. Interestingly, joint attention is also used in creating natural human-robot interaction [18]. Along the lines of head gesture recognition, detection of head nods and shakes has been found to be useful for individuals to both produce and recognize American Sign Language [6].

III. Head Movement and Dynamics Analysis

A head gesture is denoted by combinations of head motions and head fixations. This idea can be represented using dynamic and static states which represent head motions and fixations respectively. In a driving like scenario there are many spatial regions of fixations such as infotainment systems (i.e. radio, navigation system), side/rear view mirrors, and blind spots. For a given application, once fixation states of interest are determined, motion states required to go between the fixations states can be designed. In this paper we choose a few fixation states of interest in the driving scenario and corresponding motions to demonstrate proof of concept.

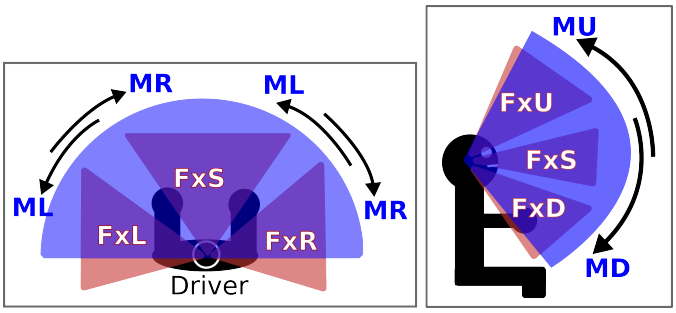

Fig. 3. Fixation states in the OHMeGA as viewed in the spatial region around the driver and motions necessary to transition between fixation states. The red triangles represent fixation regions of interest and the blue regions indicate regions of head motions.

We choose five broad fixation regions: straight fixation (FxS), right fixation (FxR), left fixation (FxL), up fixation (FxU), and down fixation (FxD). The spatial regions of these fixations as considered in this paper are shown in Fig. 3, where the left image gives a top down view of fixations in the lateral direction and the right image gives a side view of fixations in the vertical direction. The motions that allow for transitioning between these fixation states are move right (MR), move left (ML), move up (MU) and move down (MD). Using just these five fixation states and four move states, there are many possible state transitions. Fig. 4 shows the state diagram employed in OHMeGA. Notice that the state

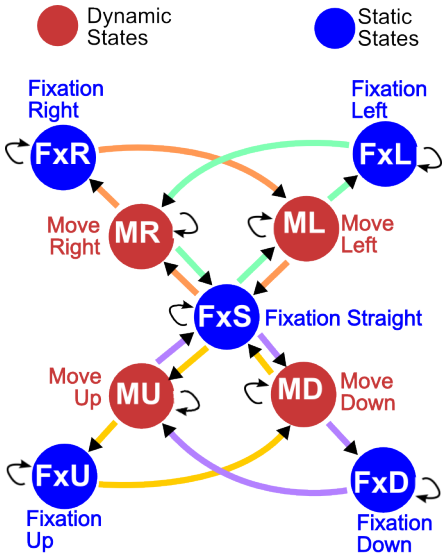

Fig. 4. State diagram of OHMeGA for head gesture analysis shows five static states (blue filled circles) representing head fixation and four dynamic states (red filled circles) representing head motion. The set of same colored arrows are used to represent one unique head gesture.

transition from one fixation state to another is restricted to go through straight fixation state. This is one of the key differences between the OHMeGA presented in [8] and the OHMeGA presented in this paper. We also assume that a particular head gesture starts from the straight fixation state. In the driving task, such assumptions are very valid since the primary function of driving is to keep focus on the road ahead. For example, in the context of lane change intent prediction [3], a driver changes the direction of his head from the road ahead towards the prospective lane a few times before changing lanes. Such head gestures are well represented using the OHMeGA analyzer.

The following subsections give details on optical flow tracking used for head motion estimation and state transitions under ideal and non-ideal conditions.

A. Optical Flow

State transitions in the OHMeGA analyzer requires estimates of the head motion. Assuming a frontal facing camera, head motions in yaw and pitch rotation angles can be approximated using horizontal and vertical motions, respectively, in the image plane. These horizontal and vertical motions are computed using Lucas-Kanade's optical flow algorithm [7]. First, interest points can be detected using well known methods like Harris corner detection and Förstner corner detection. An interest point is any distinctive point in a neighborhood that can be tracked in the consecutive frame. Optical flow vectors are then computed over sparse interest points using the equation $\mathbf{u} = -S^{-1}d$, where

$$S = \begin{bmatrix} \sum_{\mathbf{p}_i \in \aleph} I_x(\mathbf{p}_i) I_x(\mathbf{p}_i) & \sum_{\mathbf{p}_i \in \aleph} I_x(\mathbf{p}_i) I_y(\mathbf{p}_i) \\ \sum_{\mathbf{p}_i \in \aleph} I_x(\mathbf{p}_i) I_y(\mathbf{p}_i) & \sum_{\mathbf{p}_i \in \aleph} I_y(\mathbf{p}_i) I_y(\mathbf{p}_i) \end{bmatrix}$$

is a second moment matrix,

$$d = \begin{bmatrix} \sum_{\mathbf{p}_i \in \aleph} I_x(\mathbf{p}_i) I_t(\mathbf{p}_i) \\ \sum_{\mathbf{p}_i \in \aleph} I_y(\mathbf{p}_i) I_t(\mathbf{p}_i) \end{bmatrix},$$

$\aleph$ is a neighborhood of size NxN around an interest point, $I_x(\mathbf{p}_i)$ and $I_y(\mathbf{p}_i)$ are spatial gradients at point $\mathbf{p}_i \in \aleph$, and $I_t(\mathbf{p}_i)$ is the temporal gradient at point $\mathbf{p}_i \in \aleph$.

Global flow vector is then computed by a majority vote and averaged over a few frames to lessen the effects of sporadic noise. A sample at the output of optical flow tracking as applied to a video sequence containing head motions are shown in Figure 5.

B. Ideal vs. Non-ideal Conditions

Under ideal conditions, state transitions for the state diagram given in Fig. 4 is as simple as whether there is any motion and if so in what direction. We consider conditions ideal when frame rate is infinite, there is no noise in camera sensors, and all motions detected by optical flow are only due to driver intended head movements (i.e. no hand movements, vehicle vibrations, head movements due to bad road conditions). In the real world, however, the above stated conditions do not hold and the following description for state transitions takes this into account.

To demonstrate how to transition between states, we follow the green arrow shown in Fig. 4 starting at straight fixation state. An example of such a head gesture is shown in Fig. 5. In this figure, $x(t)$ and $S(t)$ represent head motion estimated by optical flow in the lateral direction and the state, respectively, at time t. Note that in Fig. 5 any head motion falling inside the noise region is taken to be zero (i.e. $x(t) = 0$ inside noise region).

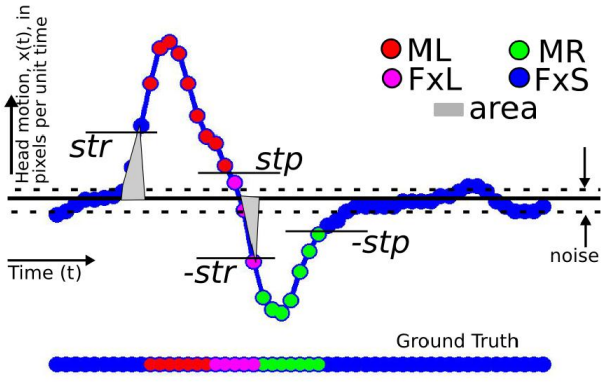

Fig. 5. Optical flow based head motion of a head gesture corresponding to the green set of colored arrows in the state diagram with corresponding ground truth. The x-axis denotes time and the y-axis denotes optical flow based head motion (pixels per unit time). Labels "str", "stp", and "area" are threshold values used in conditions for transitioning between states. Legend: red filled circles represent move left state, magenta filled circles represent left fixation state, green filled circles represent move right state, and blue filled circles represent straight fixation state. Note that the region between the dashed lines is considered to be noise.

Without loss of generality, we let the beginning of the sequence shown in Fig. 5 to be $t = 0$ and $S(0) = FxS$. In

order to transition from state FxS to state ML, $\sum_{\alpha=t_i}^{t} x(\alpha) > area$ and $x(t) > str$, where $area$ and str are threshold parameters and t_i is the time when current state $S(t)$ was entered. The state ML is retained until $x(t) < stp$ in which case it will become state FxL, where stp is a threshold value. When in state FxL, transitioning to state MR is similar to transitioning from state FxS to state ML except that the thresholds are negated and equalities are reversed: $\sum_{\alpha=t_i}^{t} x(\alpha) < -area$ and $x(t) < -str$. Once in state MR, this state is retained until $x(t) > -stp$ in which case we return to state FxS. This returns us to the starting point of the green arrows shown in Fig. 4.

The state transition given above was for a particular head gesture which has a state sequence of $\{FxS, ML, ..., ML, FxL, .., FxL, MR, ..., MR, FxS, ...\}$. Since head motions in other direction have similar optical flow output, other state transitions are readily replicable using the above mentioned state transitions.

IV. Experiments, Results and Discussion

The evaluation of our approach was conducted in two parts. The first part involves controlled laboratory setting using a driving simulator while the second part consists of the evaluation of on-road real-world driving.

A. In Laboratory

The first set of experiments was conducted in the LISA-S testbed, the setup of which is shown in Fig. 6. LISA-S testbed has the capacity to convey information to a subject using multimodal cues (i.e. visual cues using monitors and audio cues using microphones with 3D sound effects) and to synchronously collect data about the subject (i.e. eye/gaze tracking, video of the subject's head/face and the subject's foot movement). This test bed has been used for experiments in previous studies such as sequential effects on driving behavior [14] and visual attention shifts [5].

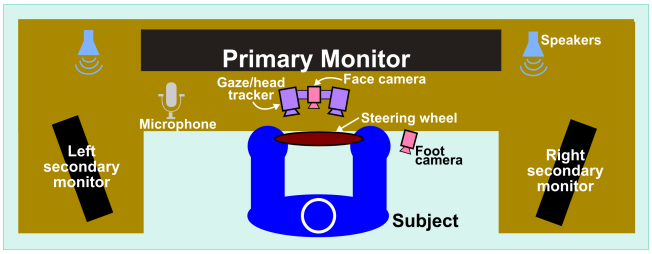

Fig. 6. Experimental setup of LISA-S test bed shows equipments used to convey information to the subject and to collect information about the subject.

In our experiment, each subject was asked to follow instructions (i.e. 'STOP' and 'GO'), either shown on the front monitor or heard using the headphones, by pressing the brake or the accelerator pedal. At the same time "distractions" in the form of simple mathematical equations were displayed on the right side monitor and the subjects were asked to answer whether the equations were correct or incorrect. Since the side monitors were placed such that the subjects had to turn

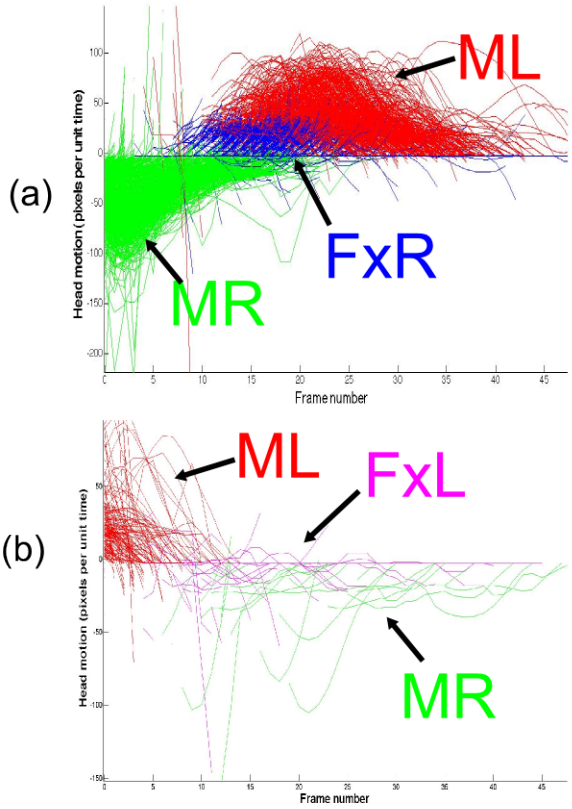

Fig. 7. Data collected using frontal facing camera from in-lab experiment is processed first using optical flow to obtain head motions, then annotated using OHMeGA analyzer and finally separated into two types of gestures. The curves in each plot above denotes optical flow head tracking and the colors represent a particular state in the OHMeGA state machine. Green lines present move right states, blue lines represent left fixation states, red lines represent move left states and magenta lines represent left fixation states.

their heads to visualize the contents, we were able to obtain desired natural head gestures.

Data from 5 subjects over 3 runs are taken from this experiment, which gave approximately 600 head gestures. The video sequence of the subject from the frontal facing camera is captured at 30 frames per second with a resolution of 480-by-704 pixels. A subject's face was approximately in the middle and occupied half of the image, as shown in the sequence of images in Fig. 2. A window, half the size of the image, centered on the image was used to find interest points for optical flow tracking.

The output of the OHMeGA analyzer as applied to the output of optical flow tracking and classified into two types of head gestures is shown in Fig. 7. Note that the head gesture represented in Fig. 7a were expected since the experiment caused the subjects to look at the right monitor. The second head gesture represented in Fig. 7b, however, was not expected. This happened mostly due to noise from hand movements on the steering wheel. One thing to note in Fig. 7b is that the magnitudes are much smaller than in Fig. 7a. In future implementations, more consideration will be placed

TABLE I

FIXATION AND MOTION STATES AS CLASSIFIED BY OHMEGA FROM IN LAB EXPERIMENT, WHERE ROWS REPRESENT THE PREDICTED STATES AND COLUMNS REPRESENT THE GROUND TRUTH.

	FxS	FxR	FxL
FxS	98.6%	26.5%	77.1%
FxR	0.9%	72.6%	20.3%
FxL	0.4%	3.5%	3.4%
Sampels	146761	3631	192

	MR	ML
MR	99.5%	4.7%
ML	0.5%	95.3%
Samples	5505	2602

TABLE II

FIXATION AND MOVE STATES AS CLASSIFIED BY OHMEGA FROM ON-ROAD EXPERIMENT, WHERE ROWS REPRESENT THE PREDICTED STATES AND COLUMNS REPRESENT THE GROUND TRUTH.

	FxS	FxR	FxL	FxD
FxS	94.3%	0%	0%	0%
FxR	0%	76.0%	0%	0%
FxL	0%	0%	83.3%	0%
FxD	0%	0%	0%	80.6%
Sampels	926	25	60	31

	MR	ML	MD	MU
MR	80.0%	0%	0%	0%
ML	0%	91.7%	0%	0%
MD	0%	0%	91.3%	0%
MU	0%	0%	0%	90.7%
Samples	145	120	69	86

on what regions to use for global flow vector calculation and in choosing the threshold values.

To evaluate the output from OHMeGA, ground truth from an independent commercial eye/head tracker, namely faceLAB, is used. Ground truth was captured in the form of head pose in the three degrees of freedom (pitch, yaw, and roll rotation angles). Taking the first derivative or the difference operation of the ground truth head pose, the resulting head motions in the yaw rotation angle were used to evaluate states in OHMeGA that are relevant to horizontal motion in the image plane. A confusion matrix from this evaluation is given in Table I. No evaluation was performed on vertical motions in the image plane because the experiment did not contain any head gestures in the pitch rotation angle.

B. From Laboratory to Roads

Although patterns of the head are similar during driving, the motion of the car and road conditions cause a lot of noise in the data. To evaluate our approach under real driving conditions, we conducted a one-subject experiment on the freeway for approximately 30-minutes at 20 frames per second. Unlike the previous experiment the "frontal-facing" camera here was a little off to the side and further from the subject, as shown in Fig. 8. Other differences in the captured images of the video sequences are: each image is 320-by-240 pixels in resolution and the subject's face is approximate $1/9$

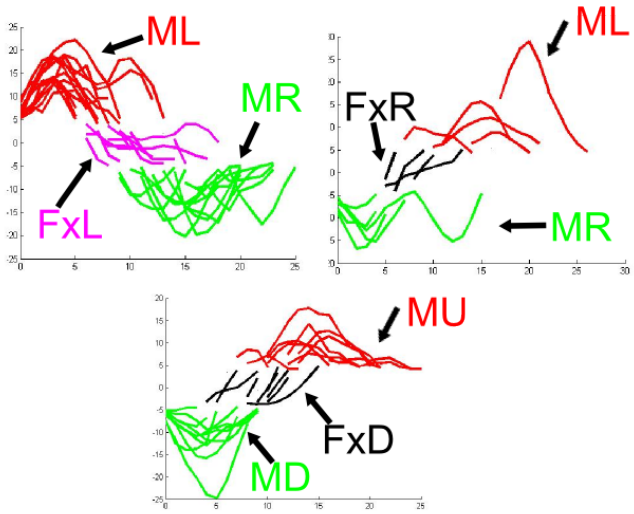

Fig. 9. Data collected using frontal facing camera from on-road experiment is processed first using optical flow to obtain head motions, then annotated using OHMeGA analyzer and finally separated into three types of gestures. The curves in each plot above denote optical flow head tracking and the colors represent a particular state in the OHMeGA state machine. Green lines present move rightdown states, black lines represent rightdown fixation states, red lines represent move leftup states, and magenta lines represent left fixation states.

the size of the image. Due to the camera distance and low resolution images, head motions detected using optical flow tracking were smaller in magnitude when compared with in lab experiments. Nonetheless many small head gestures such as glances to the side view mirrors were detected and analyzed accurately by OHMeGA.

Since vehicle motion due to road conditions caused motions in the head that were not intended by the driver, we manually selected 28 head gestures that were intended by the driver, and performed our evaluation. Ground truth used for performance evaluation was done with manual annotation. It's important to note that manual annotation is subjective, yet the performance of the OHMeGA analyzer shows promising results. Fig. 9 shows the output of OHMeGA analyzer after categorizing into three types of head gestures and Table II shows the confusion matrix of performance evaluation on this data set. In our current evaluation for vertical fixation states, we only have gestures containing down fixation states.

V. CONCLUDING REMARKS AND FUTURE DIRECTIONS

In this paper, we presented a simpler version of OHMeGA to analyze head gestures. We presented performance evaluation under two conditions: controlled laboratory experiment and uncontrolled on-road experiment. Results show an average of 97.4% accuracy in motion states for in laboratory experiment and an average of 86% accuracy overall in on-road experiment. Future work includes using face detection to indicate the region in which to perform optical flow head tracking, compensating for vehicle noise that cause unintended head motions, be able to continuously run the OHMeGA analyzer on-line, and improving state transitions from rule based to statistical learning based.

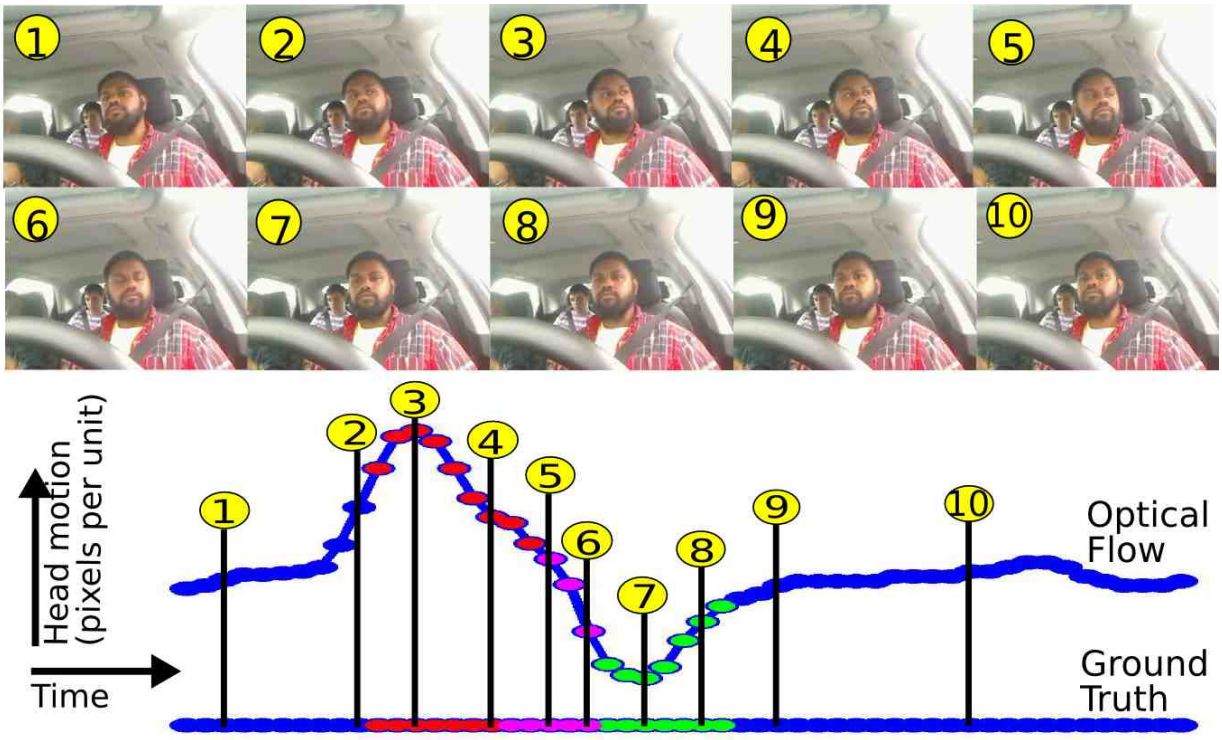

Fig. 8. Optical flow based head motion of a head gesture corresponding to the green set of colored arrows in the state diagram above with corresponding ground truth. Selected image frames are shown to understand the relationship between optical flow head tracking and visual images of the subject. The x-axis denotes time and the y-axis denotes optical flow head motion (pixels per unit time). Legend: red filled circles represent move left state, magenta filled circles represent left fixation state, green filled circles represent move right state, and blue filled circles represent straight fixation state. Note that the region between the dashed lines is considered to be noise.

REFERENCES

[1] L.M. Bergasa, J. Nuevo, M.A. Sotelo, R. Barea, and M.E. Lopez. Real-time system for monitoring driver vigilance. *Intelligent Transportation Systems, IEEE Transactions on*, 7(1):63 –77, march 2006.

[2] S.Y. Cheng and M.M. Trivedi. Turn-intent analysis using body pose for intelligent driver assistance. *Pervasive Computing, IEEE*, 5(4):28 –37, oct.-dec. 2006.

[3] A. Doshi and M.M. Trivedi. On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *Intelligent Transportation Systems, IEEE Transactions on*, 10(3):453 –462, sept. 2009.

[4] A. Doshi and M.M. Trivedi. Attention estimation by simultaneous observation of viewer and view. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 21 –27, june 2010.

[5] Anup Doshi and Mohan M. Trivedi. Head and eye gaze dynamics during visual attention shifts in complex environments. *Journal of Vision*, 12(2), February 2012.

[6] U.M. Erdem and S. Sclaroff. Automatic detection of relevant head gestures in american sign language communication. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 460 – 463 vol.1, 2002.

[7] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging Understanding Worshop*, pages 121–130, 1981.

[8] Sujitha Martin, Cuong Tran, Ashish Tawari, Jade Kwan, and Mohan Manubhai Trivedi. Optical flow based head movement and gesture analyzer (ohmega). In *Pattern Recognition (ICPR), 21st International Conference on*, Nov. 2012.

[9] B. Morris, A. Doshi, and M. Trivedi. Lane change intent prediction for driver assistance: On-road design and evaluation. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 895 –901, june 2011.

[10] E. Murphy-Chutorian and M.M. Trivedi. Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *Intelligent Transportation Systems, IEEE Transactions on*, 11(2):300 –311, june 2010.

[11] Margie Peden, Richard Scurfield, David Sleet, Dinesh Mohan, Adnan Hyder, Eva Jarawan, and Colin Mathers. World report on road traffic injury prevention. Technical report, World Health Organization, 2004.

[12] K. Sankaranarayanan, Ming-Ching Chang, and N. Krahnstoever. Tracking gaze direction from far-field surveillance cameras. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 519 –526, jan. 2011.

[13] R. Stiefelhagen, Jie Yang, and A. Waibel. Modeling people's focus of attention. In *Modelling People, 1999. Proceedings. IEEE International Workshop on*, pages 79 –86, 1999.

[14] Cuong Tran, Anup Doshi, and Mohan Manubhai Trivedi. Modeling and prediction of driver behavior by foot gesture analysis. *Computer Vision and Image Understanding*, 116(3):435 – 445, 2012. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[15] Cuong Tran and M.M. Trivedi. Driver assistance for 'keeping hands on the wheel and eyes on the road'. In *Vehicular Electronics and Safety (ICVES), 2009 IEEE International Conference on*, pages 97 –101, nov. 2009.

[16] M.M. Trivedi and S.Y. Cheng. Holistic sensing and active displays for intelligent driver support systems. *Computer*, 40(5):60 –68, may 2007.

[17] Junwen Wu and Mohan M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3):1138 – 1158, 2008.

[18] Z. Yucel, A.A. Salah, C. Merigli, and T. Mericli. Joint visual attention modeling for naturally interacting robotic agents. In *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, pages 242 –247, sept. 2009.

[19] Zhiwei Zhu and Qiang Ji. Real time and non-intrusive driver fatigue monitoring. In *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*, pages 657 – 662, oct. 2004.