Head Dynamic Analysis: A Multi-view Framework

Ashish Tawari and Moham M. Trivedi

University of California, San Diego, USA
{atawari,mtrivedi}@ucsd.edu

Abstract. Analysis of driver's head behavior is an integral part of driver monitoring system. In particular, head pose and dynamics are strong indicators of driver's focus of attention. In this paper, we present a distributed camera framework for head pose estimation with emphasis on the ability to operate reliably and continuously. To evaluate the proposed framework, we collected a novel head pose dataset of naturalistic on-road driving in urban streets and freeways. As oppose to utilizing all the data collected during the whole ride where for large portion of the time driver is front facing, we use data during particular maneuvers typically involving large head deviation from frontal pose. While this makes the dataset challenging, it provides an opportunity to evaluate algorithms during non-frontal glances which are of special interest to driver safety. We conduct a comparative study between proposed multi-view based approach and single-view based approach. Our analyses show promising results.

1 Introduction

Automatic analysis of driver behaviors is becoming an increasingly important aspect in the design of Driver Assistance System (DAS). With driver distraction and inattention being one of the prominent causes of automotive collision, we require new sensing approaches with ability to continuously infer driver's focus of attention. Eye gaze and movement are considered good measures to identify individual's focus of attention. Vision based systems provide non-contact and non-invasive solution, and are commonly used for gaze tracking. However, such systems are highly susceptible to illumination changes, particularly, in real-world driving scenario. Eye-gaze tracking methods using corneal reflection with infrared illumination have been primarily used in indoor [5] but are vulnerable to sunlight. While precise gaze direction provides useful information, coarse gaze direction, approximated by head pose and its dynamics, are often sufficient [6,3]. Head pose is a strong indicator of a driver's field-of-view and current focus of attention. It is intrinsically linked with visual gaze estimation, the ability to characterize the direction in which a person is looking. This paper presents an automatic head pose tracking system for uninterrupted driver monitoring using distributed cameras.

The two main contributions of this paper are the design of the hardware setup and annotation strategy for 'ground truth' data collection, and the development

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 536–544, 2013.

of the multi-sensory framework for improved head pose estimation. Also, important for the driving application is the design of the experiments to evaluate the realistic requirement of the vision based system in driver assistance technology. Towards this end, we gather dataset which targets spatially large head turns (away from frontal pose) since those are the times interesting events, critical to driver safety, happens. We present comparisons between single- and multi-view systems using errors statistics in yaw, pitch and roll rotation as well as the failure-rate, percentage of the time system's output is not reliable.

2 Related Studies

In driving context, challenges lie in the design and development of a system which is robust and reliable, and can operate in a 'continuous' manner. For driver monitoring task, tracking driver's head has shown much more robust performance than that of tracking eyes. In the driving distraction studies, it is suggested that non-forward glances with detectable head deviation are more severe than that without head deviation. For either prolonged or large-eccentricity, because it is more comfortable, most drivers are likely to use a combination of head and eye movements to direct their gaze to the target. It is argued that although eye-gaze measures are better, head pose still provides good estimate of driver distraction [13]. Also, in meeting room scenario, which provides more free-viewing-like opportunity, it is shown that head orientation contributes over 68% to overall gaze direction while focus of attention estimation based on head orientation alone can get 88% accuracy [11].

While there exist number of studies for head-pose estimation and tracking, their performance in real-world driving lacks proper evaluation in it's ability to operate continuously. We encourage readers to study a comprehensive survey by Murphy-Chutorian and Trivedi [8] for a good overview of different techniques and approaches for head pose estimation. In our approach, we have used facial feature and their geometric configuration along with a generic 3D face model for head pose estimation. We present some relevant literature using shape feature and geometric configuration.

Methods based on shape features analyze the geometric configuration of facial features to estimate the head orientation. Many of these existing algorithms require certain number of specific features to be visible in the image plane of the camera and use geometric constraints such as face symmetry using lip corners and both eyes [4], anatomical structure of the eye and person specific parameters [1], parallel lines between eye corners and between lip corners [12]. To avoid errors associated with precise localization of detailed facial feature, Ohue et al. proposed simple facial features - the left and right borders, and the center of the face [9]. Along with these features, the authors used a cylindrical face model to find the driver's yaw direction. Meanwhile Lee et. al. [6] used similar shape feature with ellipsoidal face model to improve the yaw estimate when the head rotates significantly.

As oppose to hand design feature and their geometric configuration, our approach can work with any set of distinct features on the face since it utilizes the

3D face model. In recent years, there has been significant advancement in the area of facial feature detection and tracking [10,14]. Their usefulness and the performance for continuous head pose estimation in an unconstrained real world setup is still not clear. It's expected that the performance would degrade as head pose deviates from frontal pose which would be the case with any system with single frontal positioned camera. Towards this end, we propose a simple distributed camera solution and conduct comparative study with single and multi-camera setup. Finally, we collect a real world head-pose dataset of naturalistic driving. Although there exist many head-pose databases ranging from images to videos, our dataset is unique with multiple cameras and consists of events with large head turns. Details of the dataset preparation including hardware setup and annotation strategy are provided in the Section 4.

3 Proposed Multi-view Framework

Spatially large head pose deviation from the frontal pose targeted in this work requires improved operating range of the head pose tracking system. For this, we use distributed cameras where each camera independently computes head pose and their results are then combined to choose the best perspective.

3.1 Appearance Based Head Pose Estimation

Pose from Orthography and Scaling (POS) [2] is used to estimate head pose from the detected facial features. POS requires at least four facial features points in general positions in the image plane and their 3D correspondences. We used generic 3D mean face model which, even though not person specific, shows promising results. For facial feature detection and tracking, we used a variant of Constraint Local Model (CLM) [10] to suit our application.

CLM utilizes parametrized shape model to capture plausible deformation of landmark locations. Using ensemble of landmark detectors, it predicts the locations of the facial landmarks. In [10], the response map of these detectors is represented non-parametrically and the landmarks' locations are optimized via subspace constrained meanshifts while enforcing their joint motion via shape model. The fitting process on an image $I^{(m,n)}$ provides a row vector $P^{(m,n)}$ for each sequence m and frame n containing $l = 66$ detected landmark positions

$$P^{(m,n)} = [x_1, y_1; x_2, y_2; \cdots x_l, y_l]$$

CLM is generally used to fit faces of various orientations, positions and sizes in the image plane. However, when streaming images from a fixed camera looking at the driver's face while driving, there are constraints in orientation, position and size of the face that can be used to further improve facial feature tracking. For each perspective, we estimate the probable face location and face size. When the tracked features are not within the expected region or if the size of the face is beyond what is normal from that perspective, the tracked facial features are rejected. Similarly, if the head pose computed from these facial features is not realizable in a driving scenario, it is also rejected.

Procedure 1. Camera Selection

Input: $prevCam$
Output: $currCam$, $prevCam$

```
Camera is numbered starting from 1 (left most) to N (right most)
i ← prevCam
if i = 0 then
    i ← INIT()
else
    if yaw_i ≥ LEFT_THR_i then
        i ← min(i − 1, 1)
    else if yaw_i < RIGHT_THR_i then
        i ← max(i + 1, N)
    end if
end if
if Camera i is active then
    Comment: A camera is active if it has measurements.
    currCam ← i
else
    currCam ← INIT()
end if
prevCam ← currCam
```

```
function INIT()
    For each camera calculate symmetry score S_i
    S_k = 0 ∀k ∈ (Set of inactive camera)
    i* ← arg max_i (S_i)
    if S_i* = 0 then
        i* = 0
    end if
    return i*
end function
```

3.2 Perspective Selection Procedure

The distributed camera setup tracks head independently in each stream and their decisions are pooled together to choose the 'best' perspective. Procedure 1 details the perspective selection criteria. During the tracking phase, the transition from one perspective to another is achieved using the thresholds in the yaw rotation angle. During initialization due to loss of track or at the very beginning, the camera is chosen based on symmetry of the face from the detected facial features. Higher symmetry-score ensures better frontal pose in the given perspective.

4 Data and Evaluation

4.1 Data Collection and Ground Truth Annotation Strategy

To evaluate our approach, we created head pose dataset consisting of multiple drivers with urban and freeway drives. The LISA-A experimental testbed, as

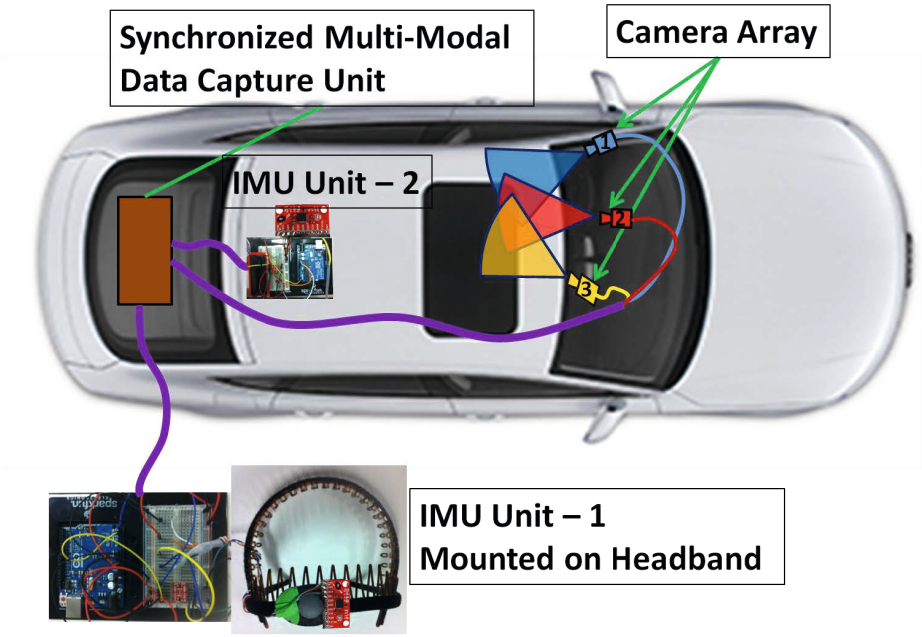


Fig. 1. LISA-A experimental testbed equipped with and capable of time synchronized capture of camera array and multiple Inertial Measurement Units (IMUs)

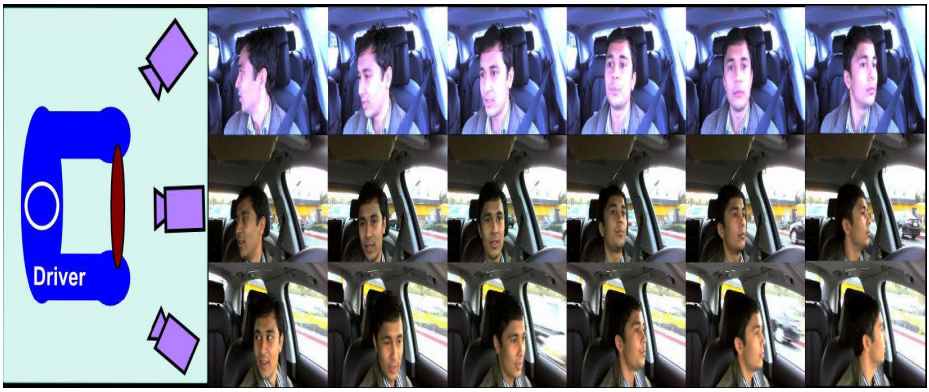
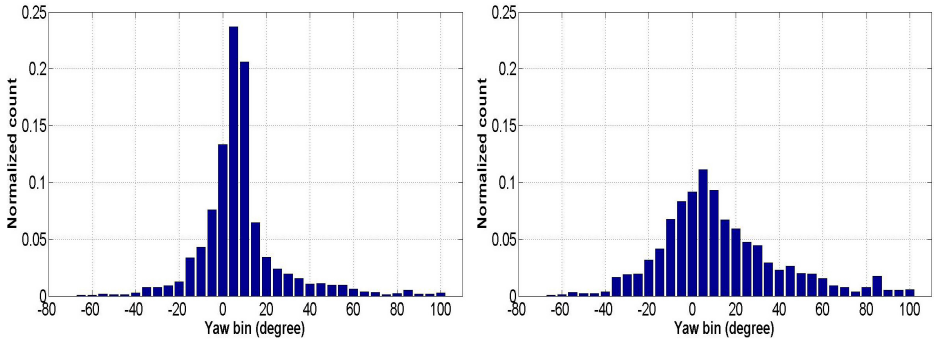


Fig. 2. An example of time-synchronized data capture from three distributed cameras

seen in Figure 1, was used to collect real-world test data. Three distributed web-cameras mounted on A-pillar, on front windshield and near rear view mirror capture face data as shown in Figure 2. These cameras provide 640x360 pixel color video stream at 30fps. In addition, the vehicle is instrumented with Inertial Motion Units (IMUs) with sensors placed on the divers head and fixed at the back of the car to track respective motion. Sensor fusion of the IMUs' data provide



(a) (b)

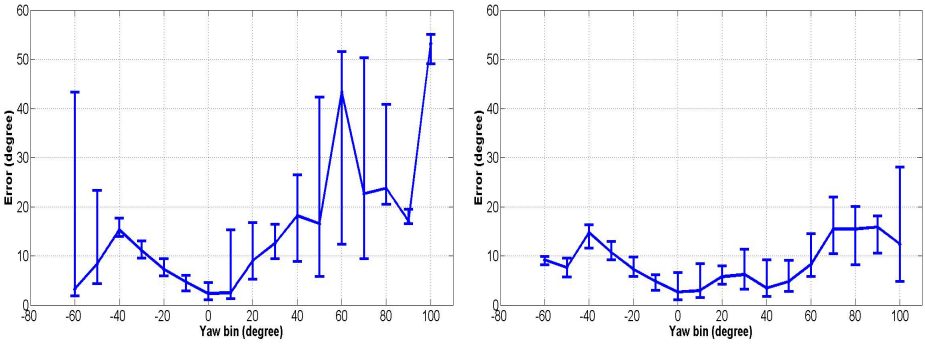
Fig. 3. Histogram of the yaw angle distribution (a) during a typical ride using all the data and (b) using select events in the dataset

Table 1. A list of events considered for evaluation, and their respective count and number of frames

Events	No. of events	Total no. of frames
Right turns	17	5062
Left turns	9	2936
Stop sign	32	7693
Right Lane change	9	1967
Left Lane Change	10	2050
Merge	5	1218
All	82	20926

precise ground truth head pose data for evaluation. Sensor fusion is required since the IMU attached to the drivers head is effected by the car movement, hence the compensation for the same is needed, which in turn, is captured by the IMU rigidly fixed to the car. The multiple IMUs involve calibrated accelerometer- and gyroscope-sensors. The IMU unit however, has some drift associated with the gyroscope, a commonly known phenomenon. This is overcome by resetting angle calculation in the beginning of each event where initial orientation is provided by hand annotation of the respective face image. Since on average each events lasts around 10 seconds the drift during this period is practically non existent.

The automobile was set up to collect data during naturalistic urban and freeway driving. Each of the two subjects drove the vehicle on similar routes through the University of California, San Diego campus in sunny weather condition causing varying lighting condition. The cameras were set to auto-gain and auto-exposure, but these adjustments have to compete with ever-shifting lighting conditions and dramatic lighting shift (e.g. sunlight diffracting around the driver) that at times saturated the camera image. All these situation are part of the evaluation, as they are typical phenomena that occur in natural driving. The placement of the cameras varies slightly for different drivers since at each



(a) (b)

Fig. 4. Absolute yaw error vs yaw angle plot. Error bar shows first quartile, median and third quartile statistics. (a) Using single camera setup and (b) using multiple camera setup.

run cameras are secured again, in case they are loosened during the drive. The drives averaged 25 minutes in duration and we analyzed different maneuvers involving large head turns as detailed in the Table 1. During urban drive, the drivers passed through many stop signs and made multiple left/right turns, and while on the freeway, they made merge, multiple lane change etc. maneuvers. Note that it's a challenging data set not just because of the real-world drive but also because data consists of and is concentrated around events with large head movement. Fig 3 shows a typical histogram of yaw angle distribution during a test ride. It can be seen that while considering the entire ride, most of the time (Fig 3(a)) drivers look front. However, the spread of the yaw angle distribution is much more for the chosen events (Fig 3(b)).

4.2 Results

We compare the performance of multi-view approach using all the three cameras with that of single-view approach with front facing camera. We, first, show in Fig 4 the absolute yaw error statistics as a function of ground-truth yaw angle with respect to front camera. The figure shows first-, second- and third quartile of the errors associated with the respective yaw bins. It can be observed that the single camera system quickly loses track with high estimation error beyond 40^0 in either direction. The multi-camera system, on the other hand, is able to keep track over much wider span with better error statistics. For quantitative evaluation over the database, two metrics, mean absolute error (MAE) and failure rate (percentage of the time system's output is not reliable), are used. Head tracking is considered lost if the estimated head pose is not available or is more than 20^0 from the ground truth in either direction for the yaw rotation angle. Number of frames, where head tracking is lost, normalized with the total number of frames over all events gives the failure rate. As shown in Table 2, failure rate of the single-view system is over 20% and that of the multi-view approach is

Table 2. Comparative evaluation of the single- and multi-view framework

Methods	Mean Absolute Error			Failure rate
	Pitch	Yaw	Roll	
Multi-view	11.5°	7.3°	4.1°	10.9%
Single-view	11.1°	8.1°	4.8°	20.7%

∼11%, a significant improvement. The MAE for pitch, yaw and roll for both approaches are comparable. This is expected since multi-camera system combines each camera independently and is bounded by the single camera accuracy.

5 Conclusion

Non-frontal glances with large head deviation are of special interest for driver safety. A Driver Assistance System (DAS) monitoring driver face/head is required to be robust and reliable in real-world conditions. Moreover, they need to perform uninterruptedly with high accuracy to be accepted and trusted by the driver. For the design and the development of such system, evaluation over real-world driving dataset is a must. There exist very few real-world naturalistic driving head pose databases. In one such database LISA-P, however, Martin et al. [7] showed using ground truth data that 95% of the time the driver's head pose was within 36^0 of forward facing in the yaw rotation angle. Therefore, there is a need for evaluating over databases more concentrated on large head movements during naturalistic on-road driving, in order to better evaluate any algorithm. Towards this end, we introduced a low cost hardware solution for ground truth data collection. Furthermore, the collected data is segmented into events with different maneuvers involving large head movement. Since a frontal single-view of the driver is insufficient for tracking head during large movements, we proposed a multi-view framework using distributed cameras. Our analysis using the collected dataset shows that the multi-view framework outperforms the single-view approach with failure rate below 11% while it's over 20% for the latter.

In future studies, we will pursue joint processing of distributed cameras for improved tracking performance. It would also be interesting to see how having more/less number of cameras than in current implementation affects the performance.

References

1. Chen, J., Ji, Q.: 3D gaze estimation with a single camera without ir illumination. In: 19th International Conference on Pattern Recognition, pp. 1–4 (December 2008)
2. Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. International Journal of Computer Vision 15, 123–141 (1995)
3. Doshi, A., Trivedi, M.M.: On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. IEEE Transactions on Intelligent Transportation Systems 10(3), 453–462 (2009)

4. Gee, A., Cipolla, R.: Determining the gaze of faces in images. Image and Vision Computing 12(10), 639–647 (1994)
5. Guestrin, E.D., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans. Biomed. Engineering 53(6), 1124–1133 (2006)
6. Lee, S.J., Jo, J., Jung, H.G., Park, K.R., Kim, J.: Real-time gaze estimator based on driver's head orientation for forward collision warning system. IEEE Transactions on Intelligent Transportation Systems 12(1), 254–267 (2011)
7. Martin, S., Tawari, A., Chutorian, E.M., Cheng, S.Y., Trivedi, M.M.: On the design and evaluation of robust head pose for visual user interfaces: Algorithms, databases, and comparisons. In: 4th ACM SIGCHI International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AUTO-UI (2012)
8. Murphy-Chutorian, E., Trivedi, M.: Head pose estimation in computer vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 607–626 (2009)
9. Ohue, K., Yamada, Y., Uozumi, S., Tokoro, S., Hattori, A., Hayashi, T.: Development of a new pre-crash safety system. In: SAE 2006 World Congress & Exhibition. SAE Technical Paper 2006-01-1461 (April 3, 2006)
10. Saragih, J., Lucey, S., Cohn, J.: Face alignment through subspace constrained mean-shifts. In: Int. Conf. on Computer Vision, pp. 1034–1041 (2009)
11. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI 2002 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2002, pp. 858–859. ACM, New York (2002), http://doi.acm.org/10.1145/506443.506634
12. Wang, J.G., Sung, E.: Em enhancement of 3D head pose estimated by point at infinity. Image and Vision Computing 25(12), 1864–1874 (2007), the age of human computer interaction
13. Zhang, H., Smith, M., Dufour, R.: A final report of safety vehicles using adaptive interface technology: Visual distraction (February 2008), http://www.volpe.dot.gov/coi/hfrsa/work/roadway/saveit/docs/visdistract.doc
14. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)