Understanding Head and Hand Activities and Coordination in Naturalistic Driving Videos
Sujitha Martin, Eshed Ohn-Bar, Ashish Tawari and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California, San Diego (http://cvrr.ucsd.edu)

Motivation
Secondary tasks performed in the vehicle tend to increase driver inattentiveness. According to a recent survey, 37% of the drivers admit to having sent or received text messages, with 18% doing so regularly while operating a vehicle [1]. Furthermore, 86% of drivers report eating or drinking (57% report doing it sometimes or often), and many reported common GPS system interaction, surfing the internet, watching a video, reading a map, or grooming.

One of the general tips for safe driving is “keeping hands on the wheels and eyes on the road.” The presence of a secondary task, however, typically violates this suggestion since it requires hand and eye coordination. While specific properties of the spatial and temporal coordination of the eye, head and hand movements are influenced by the particular tasks, it is clear that the hand usually waits for the eyes, either for target selection or for visual guidance for reaching, or both [2].

The image below illustrates a typical interaction between the driver and the infotainment system. In this particular scenario, the driver’s eyes make the first move, followed by the head and then by the hands. It shows the potential in using features representing the head and eyes in addition to hand localization as features to monitor in-vehicle driver activities.

Research Objective
In this work, we develop a fusion framework for activity analysis using head, eye, and hand cues. The framework is used in order to detect three important recurrent driver activities of interest: interaction with the steering wheel, the gear and the instrument cluster.

References

Activity Analysis: Integration of head and hand cues

- **Hand cue-based activity recognition**: The proposed framework is purely vision-based, with no markers or intrusive devices. Hand activity is detected in each zone using a visual descriptor (pyramidal HOG) and a multiclass SVM classifier. These cues were shown to work well against a variety of other visual cues [3]. A weight is jointly learned for each class in order to leverage the structure among different activity regions.

  \[ r^2 = \text{pyramidal HOG} \]

- **Head cues**: Facial features-based geometric approach is used for head pose (pitch, yaw, roll) estimation and the level of eye opening. The eye opening at time \( t \) is defined as follows,

  \[ e(t) = A_{\text{head}}(t) \]

  where, \( A_{\text{head}}(t) \) is the area of the convex hull of the facial landmarks and \( e_{\text{head}} \) is a normalization constant learned for each driver to represent his or her normal eye opening state. The plots above illustrate the head cues with the mean (solid line) and standard deviation (semi-transparent shades) of two features (pitch and eye opening) for three different activities, using the naturalistic driving dataset.

- **Hand and Head cues-based Activity Recognition**: First, the hand cues are summarized using normalized scores, outputted by the hand-only learned SVM.

  \[ p(h) = \frac{\text{HOG}(w, x)}{\sum \text{HOG}(w, x)} \]

These are abbreviated as \( p_h \) and calculated at every frame. Next, the hand-based activity classification is rescored using the temporal head and eye cues. The integrated feature vector

\[ \psi(t) = \begin{pmatrix} p_h(t) \\ e(t) \\ \text{IC}(t) \end{pmatrix} \]

is given to a hierarchical multiclass SVM to produce the final hand, head, and eye-based activity classification.

Results and Discussion

- **The naturalistic driving dataset** is collected with multiple drivers (three male and one female) of varying ethnicity and varying age from 20 to 30, as well as showing features representing the head and eyes in addition to hand localization as features to monitor in-vehicle driver activities.

- **A total of 11,147 frames were annotated**: 7429 frames of two hands in the wheel region, 679 frames of hands on the gear, and 3039 frames of interaction in the instrument cluster region.

- **In order to capture head and hand cue dynamics**, head and eye cues are calculated over a temporal window. The effect of changing the time window is shown below. We notice how increasing the window size up to two seconds prior to the onset of activity improves performance. With larger window, the cue becomes less discriminative.

- **The confusion matrices, shown below**, validate the integration of head and hand cues for improved activity recognition. Note that, with hand-cues only, instrument cluster (IC) and gear classification are often confused. One reason being the presence of arm in the gear region while interaction occurs with the IC.

- **Below is a visualization of the advantage in integrating head and hand cues for driver activity recognition**. Note how the incorrect hand-based predictions were corrected by the rescoring based on head and head cues.

- **Future work would extend the activity grammar to include additional activities of more intricate maneuvers and driver gestures. Furthermore, combining the head pose with the hand configuration to produce semantic activities can be pursued using temporal states models.**

![Image of driver activity classification](Image)

![Image of effects of varying time window](Image)

![Image of hand cue-based activity recognition](Image)