Estimating attention level

Literature review

# 

# Presenting the literature

Driver drowsiness has been identified as a key factor in causing road accidents. To address this issue, the need for an effective driver drowsiness detection system, capable of alerting drivers prior to an accident, has been highlighted by research. The main idea is to find a way to measure attention: a low level of attention in the driver, must lead to the activation of a correction feedback.

According to [1] drowsiness can be determined by three measures: vehicle-based measures, behavioral measures and physiological measures.

The behavioral measures include yawning, eye closure, eye blinking and head pose. Some support is, indeed, given by [2]: they established a significant relationship between eye blinking and attention and engagement. It demonstrates that viewers modulate their blink rates as a function of their engagement with visual content. Specifically, it shows that individuals tend to blink less when viewing content they perceive as more relevant or important, suggesting that the inhibition of blinking is directly tied to moments when viewers are most engaged.

For this reason, extensive research has been carried out to develop a camera-based system capable of assessing attention levels through the analysis of eye blinking patterns.

[3] focuses on developing a real-time algorithm to detect eye blinks from video sequences using standard cameras. The algorithm capitalizes on the advancements in landmark detectors, trained on “in-the-wild datasets” that have shown robustness against variations in head orientation, lighting, and facial expressions [4]. These detectors are capable of accurately identifying eye landmarks, allowing for the reliable estimation of eye opening levels through a metric called the Eye Aspect Ratio (EAR).

The Eye Aspect Ratio (EAR) is a scalar value derived from the facial landmarks around the eye, which quantifies the eye's openness. It's calculated using the vertical and horizontal distances between specific landmarks on the eye. The formula is given as:



In this formula, P1, P2, P3, P4, P5 and P6 represent the 2D positions of the six key landmarks around the eye. P1 and P4 are points at the corners of the eye, forming the horizontal line. P2 and P5 are at the top, and P3 and P6 are at the bottom of the eye, forming two vertical lines. The EAR essentially measures the ratio of the distances between the vertical landmarks to the distance between the horizontal landmarks. This ratio decreases as the eye closes.

They describe eye blinking as a dynamic process characterized by variations in speed, eye squeezing degree, and duration, typically lasting between 100-400 milliseconds.

Recognizing that per-frame EAR values alone might not accurately identify blinks (due to other factors like prolonged eye closures or facial expressions), they propose to use a classifier. This classifier considers a sequence of EAR values across a temporal window rather than relying on a single frame's measurement. Specifically, they suggest using a linear Support Vector Machine (SVM) classifier trained on patterns of blinking and non-blinking EAR sequences. The SVM classifier is designed to recognize the distinct temporal pattern of an eye blink in a sequence of EAR values spanning several frames, thereby improving the reliability of blink detection.

This method addresses the challenge of distinguishing genuine blinks from other eye closures by incorporating the temporal dynamics of blinking.

The paper [5] presents a novel approach to measure a person’s level of attention using basic equipment like a standard camera. This method analyzes eye blinks and head movements by employing facial landmarks: fewer blinks and lesser head movement signify higher attention.

They introduced a formula to combine the normalized blink rate and head movement data to estimate the overall level of attention. This formula considers the normalized number of blinks and the extent of head movements to provide a composite score indicating the attention level. The resulting score ranges between 0 and 1, where higher scores suggest higher levels of attention.

Another approach was proposed by [6]. It introduced a cutting-edge Driver Monitoring System (DMS) aimed at enhancing driver safety by monitoring behavior through facial landmark estimation. This system utilizes an infrared (IR) camera to capture video data for analyzing the driver's head posture and eye area. The use of IR cameras addresses common challenges related to variable lighting conditions, ensuring the system's reliability across different driving environments.

The system comprises two modules, each designed to recognize specific driver behaviors. The first module evaluates the driver’s head movements to detect inattention, while the second uses an eye-closure recognition filter to identify drowsiness based on the continuity of eye closures.

To identify the driver’s inattention, the system initiates the head pose estimation process. Head pose estimation provides crucial information about the driver’s head angle and gaze direction: they looked for signs of drowsy driving, such as shaking the head from side to side or consistent nodding, as well as indications of the driver’s focus on the road ahead based on the orientation of the head when the steering wheel is fixed.

[7] proposed a different kind of index for measuring attention: the PERCLOS. It is a pivotal metric for measuring drowsiness. This metric quantifies the extent of eye closure over a specific time period.

PERCLOS = 100 x [(number of eye closure events)/(total number of frames)]

A higher PERCLOS percentage indicates a greater proportion of time spent with closed or partially closed eyes, suggesting increased drowsiness. Typically, it is required to define a threshold above which it is considered that the driver is experiencing drowsiness, some typical values are 20% or 40%.

# 

# 

# 

# 

# Datasets

From [8], underlined are available:

* **300W-LP**: the 300W-LP (Large Pose) is a synthetic extension of the 300W database, generated to augment the number of challenging samples with extreme poses. It includes 122 450 images with a yaw angle in range ±89 degrees.
* **AFLW**: Annotated Facial Landmark in the Wild is a challenging dataset which was collected from the internet, in totally unconstrained conditions. It contains a collection of 25, 993 faces with head poses ranging between ± 120◦ for yaw and ± 90◦ for pitch and roll. The pitch, yaw and roll angles were obtained automatically from the labeled landmarks using the POSIT algorithm, assuming the structure of a mean 3D face, for this reason, several annotation errors were found.
* **AFLW2000**-**3D**: This dataset contains the first 2000 identities of the in-the-wild AFLW dataset which have been re-annotated with 68 3D landmarks using a 3D model which is fit to each face. Consequently, this dataset contains accurate fine-grained pose annotations and is a prime candidate to be used as a test in head pose estimation task. Yaw varies ±120◦, while roll and pitch ±90◦ .
* **AFW**: Annotated Faces in the Wild represents a small database (it’s a subset of AFLW), which is normally used for testing purposes only. AFW has 250 images and inside these images 468 faces in a very challenging environment are included. The yaw angles vary between ± 90◦ with a step size of 15◦. The ground-truth is manually annotated, so it may contain errors.
* **AISL**: The Aisl head orientation database is a collection of small scale head images with various backgrounds of an indoor scene. This dataset contains 6480 images of 20 subjects under 36 yaw angles, 3 pitch angles and 3 different backgrounds. The orientation is determined by two categories: yaw angle in 360◦ with an interval of 10◦, and pitch angle in the range ±45◦ with an interval of 45◦ .
* **AutoPOSE**: It’s a large-scale dataset that provides 1.1 million images taken from a car’s dashboard view. AutoPOSE’s ground-truth head orientation was acquired with a sub-millimetre accurate motion capturing system placed in a car simulator. The rotations are limited to the range [– 90◦, + 90◦], the average pitch angle is shifted in the negative values of the rotation angles, this is due to the placement of the camera in the dashboard.
* **BioVid Heat Pain**: It contains videos and physiological data of 90 persons subjected to well-defined pain stimuli of 4 intensities, built for the development of automatic pain monitoring systems. It includes information about head pose of the recorded subjects for all 3 angles pitch, yaw, roll, all in the range ±50◦ .
* **BIWI Kinect**: It’s gathered in a laboratory setting by recording RGB-D video of different subjects across different head poses, using a Kinect v2 device. It contains roughly 15, 000 frames and the rotations are ±75◦ for yaw, ±60◦ for pitch and ±50◦ for roll. A 3D model was fit to each individual’s point cloud and the head rotations were tracked to produce the pose annotations. This dataset is commonly used as a benchmark for pose estimation using depth methods that attests to the precision of its labels.
* **BJUT-3D**: The database consists of 46 500 images collected from the 3D faces of 250 male and 250 female participants. The total number of poses in the database is 93. The pitch rotation is quantized into 9 angles [– 40◦, +40◦], where the diference between two consecutive poses is 10◦ . Similarly, the yaw rotation is divided into 13 angles [-60◦, +60◦], with the same angular step size as for the pitch.
* **Bosphorus**: It contains 5 thousand high resolution face scans from 105 different subjects. The 3D scans are obtained by a commercial structured-light based 3D digitizer. It offers 13 discrete head pose annotations (seven yaw angles, four pitch angles, and two roll angles), with different facial expressions and occlusions.
* **BU**: The Boston University Head Tracking dataset includes only 200 images and 5 subjects, which is the main drawback of this database. The acquisition process is repeated in two sessions: initially illumination conditions are uniform; then subject faces are exposed to rather complex scenarios with changing illumination. All three rotation angles were recorded thanks to a magnetic tracker attached to each participant’s head. Pose variation is mainly less than 30◦. Since the presence of facial occlusions (eyeglasses, facial hair, etc.) is very limited, most methods perform very well.
* **CAS-PEAL**: The CAS-PEAL is a large dataset having 99 594 images, with a total number of 1040 participants, with 595 males and 445 female subjects. The CAS-PEAL dataset contains a total of 21 poses combining diferent yaw and pitch angles: the yaw orientation varies between – 45◦ and + 45◦ with an interval of 15◦ between two consecutive poses; the pitch orientation has only three poses – 30◦, 0◦, and + 30◦. Although the dataset has sufficient data for evaluation and training, its complexity is low, as the number of poses is quite limited.
* **CAVE**: The Columbia Gaze dataset contains a total of 5880 images of 56 different subjects (32 male, 24 female) of different ethnic groups and ages. The dataset is mainly created to solve the gaze estimation task, but contains also information about head pose of the participants, therefore it can be used to solve the discrete head pose estimation task. For each subject a combination of fve horizontal head poses (0◦, ± 15◦, ± 30◦), seven horizontal gaze directions (0◦, ± 5◦, ± 10◦, ± 15◦), and three vertical gaze directions (0◦, ±10◦) are available.
* **CCNU**: All images in CCNU are low-resolution images collected in a classroom. The database consists of 58 participants, captured in 75 diferent poses, for a total number of 4 350 images. The face images are collected so that illumination conditions and facial expressions are changing, thus adding more complexity to the images. For obtaining the ground-truth data, SensoMotoric Instruments (SMI) eye tracking glasses are used. The head orientation changes from – 90◦ to + 90◦ in the horizontal direction, while the vertical direction spans in the range – 45◦ to + 90◦ .
* **CMU Multi-Pie**: This is a database collected from subjects exhibiting multiple expressions under different illumination conditions in a constraint environment. All high-resolution images are captured using a system of 15 cameras for a total of 75 thousand images. The only angle of rotation available is the yaw with an incrementation step of 15◦ .
* **CMU Panoptic Dataset**: It’s a large scale dataset providing 3D pose annotations for multiple people engaging in social activities. It contains 65 videos with multi-view annotations captured inside a dome from approximately 30 HD cameras. The panoptic dataset includes 3D facial landmarks and calibrated camera extrinsics and intrinsics, but does not include head pose information. Using landmarks and camera calibrations it is possible to locate and crop images of subjects’ heads and compute the corresponding camera-relative Euler angles. After processing the dataset to address the head pose problem, it contains 1,342,018 images. The yaw angle distribution is almost uniform and ranges in ±179◦, but at angles near 90◦ and – 90◦ there are fewer images due to the effect of Gimbal lock. For the two angles pitch and roll the magnitudes are in the range ± 89◦ .
* **CMU-PIE**: The CMU Pose, Illumination, and Expression (PIE) dataset contains over 40,000 facial images of 68 people. Using the CMU 3D Room each person is imaged across 13 different poses, under 43 different illumination conditions and with 4 different expressions. The pose ground-truth was obtained with a 13 cameras array, each positioned to provide a specific relative pose angle. This consisted of 9 cameras at approximately 22.5◦ intervals across yaw, one camera above the center, one camera below the center, and one in each corner of the room.
* **DAD-3DHeads**: This is an in-the-wild database that contains a variety of extreme poses, facial expressions, challenging illuminations, and severe occlusions cases. It consists of 44 thousand images annotated using a 3D head model, a non-linear optimization algorithm and a final manual adjustment. To validate head pose annotations the rotation matrices were compared to the groundtruth matrices from the BIWI dataset.
* **Dali3DHP**: This database is an extreme head pose database collected from a camera mounted on a treadmill. The dataset was collected in two different sessions from 33 individuals. Ground-truth data is collected using Shimmer sensor 2 which is attached to each person’s head. The database is large since it contains more than 60,000 depth and color images. All the three rotation angles pitch, yaw and roll were defined at the time the acquisition took place, covering the following head angles: pitch [− 65.76◦, + 52.60◦], roll [−29.85◦, + 27.09◦], and yaw [− 89.29◦, + 75.57◦ ].
* **DD-Pose**: It contains 330 thousand measurements from multiple cameras acquired by an in-car setup during naturalistic drives by 27 subjects. Large out-of-plane head rotations and occlusions are induced by complex driving scenarios, such as parking and driver-pedestrian interactions. Precise continuous 6 DoF head pose annotations are obtained by a motion capture sensor and a novel calibration device. The angles vary in the following ranges, ignoring outliers with less than 10 measurements in a 3◦ neighborhood: pitch ∈ [– 69◦, + 57◦], yaw ∈ [– 138◦, + 126◦], roll ∈ [– 63◦, + 60◦ ].
* **DriveAHead**: It’s another driver head pose dataset, it contains frame-by-frame head pose labels obtained from a motion-capture system for 20 subjects (about 1 million of frames). It includes parking maneuvers, driving on the highway and through a small town, different occlusions and illuminations, thus providing distributions of head orientation angles and head positions which are typical for naturalistic drives. Images were collected with a resolution of 512×424 pixels, 6 DoF, the range of angles is [– 45◦, + 45◦] for pitch, [– 40◦, + 40◦] for roll and mainly [– 90◦, + 90◦] for yaw.
* **ETH**: The ETH Face Pose Range Image Dataset contains more than 10 thousand images of 20 persons (3 of them being female) at a resolution of 640 × 480 pixels. Each person freely turned her head while the scanner captured range images at 28 fps. Yaw varies between -90◦ to + 90◦ , pitch between – 45◦ to +45◦ , whereas roll is not considered.
* **FacePix**: The FacePix database is built depicting 30 individuals, for a total number of 5 430 images. It is an imbalanced dataset with 25 males and 5 females. Yaw rotation varies from – 90◦ (extreme left profile) to + 90◦ (extreme right profile), with a step size of 2◦; no other rotation angles were considered.
* **GI4E-HP**: It contains 36 thousand images from 10 subjects recorded with a web-cam in an in-laboratory environment. Head pose annotations are given in 6 DoF using a magnetic reference sensor. All transformations and camera intrinsics are provided. Head pose annotations are given relative to an initial subjective frontal pose of the subject.
* **GOTCHA-I**: This dataset is a collection of 682 videos of 62 subjects in 11 different indoor and outdoor environments to address both security and surveillance problems. To obtain ground-truth a 3D head model is reconstructed and elaborated using Blender software. There are 137, 826 labeled frames with 2223 head pose per subject in the range of [– 40◦, + 40◦] in yaw, [-30◦, +30◦] in pitch and [– 20◦, + 20◦] in roll, with a step of 5◦ .
* **ICT-3DH**P: It’s a large dataset which was collected in-the-wild, i.e. captured in an unconstrained environment. All images were acquired through the Polhemus Fastrack1 fock of birds tracker attached to a cap of the participants that contains a magnetic sensor, so that the dataset contains both RGB and depth data. The database is evaluated for all three rotation angles including pitch, yaw and roll. No accurate information about the angle ranges is provided.
* **IDIAP Head Pose**: It contains 66, 295 head images stemming from a 8 video meeting recording, each approximately one minute in duration, of a few people in a meeting room. In each sequence, two subjects, which are always visible, were continuously annotated using a magnetic sensor. Therefore, each image has a complete annotation of a head pose orientation from pitch (range[– 60◦, + 15◦ ]), yaw (range ± 60◦) and roll (range ± 30◦) angles.
* **M2FPA**: This dataset totally involves 397, 544 images of 229 subjects with 62 poses (including 13 yaw angles, 6 pitch angles and 44 yaw-pitch angles), 4 attributes and 7 illuminations. There are 6 classes for pitch in the range of [– 30◦, +45◦] with a step increment of 15◦ and 13 measurements for yaw in the range ±90◦ with a step increment of 15◦ .
* **McGill**: The database consists of 60 videos of 60 different participants, in total it contains 18, 000 video frames. The videos were recorded in both indoor and outdoor environments. The participants were free to behave as they want during the video collection process, therefore arbitrary illumination conditions and background clutter are present, especially outdoors. Only yaw angles are estimated using a semi-automatic procedure, with variation in the range [– 90◦, + 90◦ ].
* **MDM corpus**: The Multimodal Driver Monitoring database was collected with 59 subjects recorded while driving a car and performing various tasks. To record the head pose the Fi-Cap device was used, this continuously tracks the head movement of the driver using fiducial markers, providing frame-based annotations to train head pose algorithms in naturalistic driving conditions. This set consists of 48.9 h of recordings (10, 541, 166 frames), it covers a large range of head poses along all three rotation axes due to the large number of subjects included, and the variety of primary and secondary driving activities considered during the data acquisition. Yaw angles range around the origin spanning between – 80◦ to 80◦, pitch angles have an asymmetric range spanning from – 50◦ to 100◦ .
* **MTFL**: The Multi-Task Facial Landmark dataset contains 12, 995 outdoor face images from the web. These images are from CUHK Face Alignment database and AFLW dataset. Each image is annotated with a bounding box and five facial landmarks. There are ground-truth annotations for gender, age, smiling, wearing glasses and head pose. For the latter, the images are manually categorized in 5 discrete classes: Left-profile, Left, Frontal, Right, Right-profile.
* **Pandora**: It has been specifically created for head center localization, head pose and shoulder pose estimation and is inspired by the automotive context. A frontal fixed device acquires the upper body part of the subjects, simulating the point of view of the camera placed inside the dashboard. Subjects also perform driving-like actions, such as grasping the steering wheel, looking to the rear-view or lateral mirrors, shifting gears and so on. Pandora contains more than 250 thousand full resolution RGB (1920× 1080 pixels) and depth images (512 × 424) acquired with a Microsoft Kinect 1 device. Subjects perform wide head movements: ± 70◦ roll, ± 100◦ pitch and ± 125◦ yaw. Garments as well as various objects are worn or used by the subjects to create head occlusions. The ground-truth annotations have been collected using a wearable Inertial Measurement Unit (IMU) sensor.
* **Pointing’04**: It is one of the oldest databases, released in 2004, which was considered as the classical benchmark for HPE (in some studies is also called PRIMA database). Despite its age, it’s still used for research purposes, due to its challenging nature and a large variety in consecutive poses [29–32]. A total number of 15 participants (between 15 and 40 years) were involved for image acquisitions. Some of them wear eyeglasses or show facial hairs, thus increasing the task complexity. Images were collected in an indoor lab environment, with very low illumination conditions. Each participant is asked to look at some markers on the wall, and two rotation angles (yaw and pitch) are annotated through a subsequent manual labeling process (thus introducing some errors). The head orientation varies between ± 90◦ both in the horizontal and vertical directions, while the difference between two consecutive poses in horizontal and vertical orientation is kept at 15◦ and 30◦, respectively.
* **SASE**: This is a 3D database collected through Kinect 2 camera. It consists of both RGB and depth images of 32 male and 18 female subjects. The total number of frames is 30, 000. All subjects have different ethnicity and hairstyles, with an age range of 7–35 years. All three rotation angles pitch, yaw, and roll are considered. All participants have different facial expressions during image acquisition, so that, along with head pose estimation, the database may also be used for emotion recognition. For each person a large sample of head poses are included, within the bounds of yaw from – 45◦ to 45◦ , pitch – 75◦ to 75◦ and roll – 45◦ to 45◦ of rotation around each axis.
* **SyLaHP**: The Synthetic dataset for Landmark based Head Pose estimation was proposed by Werner et al. along with a benchmark protocol to learn head pose on top of any landmark detector (called HPFL). It contains about 101 thousand synthetic images from 30 subjects, with varying ethnicity, age and gender. The angles are in the ranges: ± 70◦ for pitch, ± 90◦ for yaw and ±55◦ for roll.
* **SynHead**: This is a large-scale synthetic dataset for head pose estimation in videos containing 10 head models (5 female and 5 male), 70 motion tracks and 510 960 frames. Such synthetic dataset, which considers all Euler angles, generates 100% reliable ground-truth to compensate for errors existing in manually annotated datasets. The Euler angles are in the range of [– 100◦, +100◦ ].
* **Synthetic**: The Synthetic image database is a large database of 74, 000 high quality images taken from head models. A total of 37 sequences have been considered, where each sequence includes 2000 frames. The head pose in face images covers ± 50◦ of roll, ± 75◦ for yaw, and ± 60◦ for pitch. The database is quite challenging as different ages, races, and facial expressions are included.
* **Taiwan RoboticsLab**: It contains 6660 images of 90 subjects. For each subject there are 74 images, where 37 images were taken every 5 degrees from right profile (defined as + 90◦) to left profile (defined as – 90◦) in the yaw rotation using camera array and the remaining 37 images were generated (synthesized) by the existing 37 images using commercial image processing software in the way of flipping them horizontally.
* **UbiPose**: This dataset relies on videos from the UBImpressed dataset, which has been captured to study the performance of students from the hospitality industry at their workplace. The data are recorded using a Kinect 2 sensor, however the ground-truth head pose is indirectly inferred from facial landmarks. The validated inferred head poses are 10.4 thousand, most frames fall within a [20◦, 40◦] interval.
* **UET-Headpose**: The UET-Headpose dataset was created to capture the head pose of annotated people in many conditions, it includes 12, 848 images obtained from 9 people. The dataset has a uniform yaw angle distribution for all directions in the range [– 179◦, 179◦]. The dataset is obtained by having the annotated people rotated all yaw directions when collecting the dataset. Therefore, it is possible to learn all yaw angles within a 360◦ range.
* **UMD Faces**: This dataset has 367, 888 annotated faces of 8277 subjects. It contains information about bounding boxes (verified by humans), twenty-one keypoint locations, Euler angles and the gender of the subject. These annotations have been generated using the All-in-one CNN model, therefore the dataset may contain erroneous annotations, especially for the pitch, yaw and roll angles.
* **VGGFace2**: This is a very large HPE database which has been released in 2018. It contains 3.31 million images. The total number of participants to create this content is 9131, whereas the average number of images per subject is 362. The database is constructed with images downloaded from Google Image Search and shows large variations in pose, illumination, age, profession, and ethnicity. However, pose (pitch, yaw and roll) is estimated using pre-trained pose classifiers defining 5 classes for angles in ranges [– 100◦, – 40◦), [– 40◦, – 10◦), [– 10◦, + 10◦), [+ 10◦, +40◦) and [+ 40◦, + 100◦ ).

# 

# 

# 

# 

# 

# 

# References

[1]: [Detecting Driver Drowsiness Based on Sensors: A Review Arun Sahayadhas \*, Kenneth Sundaraj and Murugappan Murugappan](https://www.researchgate.net/publication/233881105_Detecting_Driver_Drowsiness_Based_on_Sensors_A_Review)

[2]: [Blink Rate Patterns Provide a Reliable Measure of Individual Engagement with Scene Content. Carolyn Ranti, Warren Jones, Ami Klin & Sarah Shultz](https://www.nature.com/articles/s41598-020-64999-x)

[3]: [Real-Time Eye Blink Detection using Facial Landmarks. Tereza Soukupova and Jan Čech](https://vision.fe.uni-lj.si/cvww2016/proceedings/papers/05.pdf)

[4]: [Incremental Face Alignment in the Wild. Akshay Asthana1 Stefanos Zafeiriou1 Shiyang Cheng1 Maja Pantic1,2 1Department of Computing, Imperial College London, United Kingdom 2EEMCS, University of Twente, Netherlands](https://ibug.doc.ic.ac.uk/media/uploads/aasthanacvpr2014.pdf)

[5]: [Estimating Attention Level from Blinks and Head Movement. Masaaki Goto, Tetsuo Tanaka, and Kazunori Matsumoto Kanagawa Institute of Technology, Japan](https://easychair.org/publications/open/3fK6)

[6]: [Real-time driver monitoring system with facial landmark-based eye closure detection and head pose recognition. Dohun Kim, Hyukjin Park, Tonghyun Kim, Wonjong Kim, Joonki Paik](https://www.nature.com/articles/s41598-023-44955-1)

[7]: [Feeling Drowsy? An In-Depth Look at PERCLOS for Driver Drowsiness Detection. The RidgeRun.ai Team](https://ridgerunai.medium.com/driver-drowsiness-detection-using-perclos-3cacd08e2359)

[8] [Deep Learning for Head Pose Estimation: A Survey. Andrea Asperti, Daniele Filippini](https://www.researchgate.net/publication/370288280_Deep_Learning_for_Head_Pose_Estimation_A_Survey)