Geometric Head Pitch Estimation from Thermal Facial Landmarks Using a Multitask Convolutional Neural Network

Estimación Geométrica del Cabeceo en Imágenes Térmicas Mediante Landmarks Faciales y una Red Neuronal Convolucional Multitarea.

Salvador Calderon Uribe[1], Uriel Calderon Uribe[2], Irving Armando Cruz Albarran[1]
[1]Laboratorio de Visión Artificial y Termografía/Mecatrónica Facultad de Ingeniería, Universidad Autónoma de Querétaro, Campus San Juan del Río, San Juan del Río, CP. 76807, México. scalderon02@alumnos.uaq.mx, irving.cruz@uaq.mx
[2]Laboratorio de Visión por Computadora, Tecnológico Nacional de México/ITS Sur de Guanajuato, Uriangato 38980, México. u.calderon@itsur.edu.mx

Abstract

Monitoring driver visual behavior provides valuable information about the driver's level of attention and plays an important role in developing intelligent driver assistance systems. This paper presents a multitask Deep learning model that uses thermal facial images to estimate facial landmarks and classify overall head orientation (frontal and lateral) simultaneously. The predicted outputs are then used to compute two sets of geometric descriptors for estimating head pitch. The proposed model was evaluated on the Charlotte-Thermal Face dataset, achieving a mean absolute error (MAE) of approximately 9 pixels for facial landmark localization and 87% classification accuracy for head orientation estimation. Furthermore, the proposed descriptors exhibited consistent behavior across different head inclinations, providing complementary geometric information that can be incorporated as an additional feature in driver monitoring systems.

Palabras clave: Thermography; Facial Landmarks; Face Orientation; Multitask Model; Pitch Estimation.

Introduction

Recent advances in computer vision and Advanced Driver Assistance Systems (ADAS) have significantly improved with features like collision avoidance, lane-keeping assistance, and automatic emergency braking (Jha *et al.,* 2023; Kim *et al.,* 2023). However, the driver's physical and cognitive condition remains critical to ensuring safe vehicle operation because the driver retains ultimate responsibility for vehicle control. Consequently, fatigue, distraction, and alcohol impairment continue to be major contributors to traffic accidents (Kim *et al.,* 2023). According to the National Institute of Public Health (INSP), road traffic accidents account for approximately 24,000 fatalities annually in Mexico, ranking the country seventh worldwide and third in Latin America in traffic-related deaths (Caminos y Puentes Federales, 2022). Similarly, the Mexican Institute of Transportation (IMT) identifies driver inattention as a leading cause of road accidents (Caminos y Puentes Federales, 2022).

In this context, there is an increasing need for intelligent driver monitoring systems that can continuously assess drivers' attention states. Among different behavioral cues, head orientation has emerged as a valuable feature because it provides information about the driver's gaze direction and potential signs of fatigue or visual distraction (Dong *et al.,* 2010). Several approaches have been proposed in the literature to assess the driver's condition and attention level, including the analysis of physiological signals (Qiu *et al.,* 2019), the detection of cognitive distractions (Li & Busso, 2014), and the recognition of emotional states (Abdic *et al.,* 2016). Among the various factors associated with unsafe driving, visual distraction caused by looking away from the road is recognized as one of the leading risk factors contributing to traffic accidents (Dingus *et al.,* 2006). Consequently, head pose estimation has emerged as an effective strategy for inferring the driver's focus of attention, since head orientation provides valuable information about gaze direction and changes in alertness. However, estimating head pose from RGB imagery remains challenging due to variations in illumination, shadows, and cluttered backgrounds, which can reduce the reliability of vision-based systems (Kim *et al.,* 2023). As an alternative, thermal infrared imaging combined with computer vision and deep learning techniques has attracted increasing attention because it is considerably less sensitive to ambient lighting conditions while preserving facial structural information. Consequently, several studies have explored using

thermal imaging for robust head orientation estimation under varying illumination conditions. For instance, (Kim *et al.,* 2023) developed a real-time driver monitoring system which estimates facial landmarks from thermal images in order to extract head pose and eye region information for assessing driver attention. Similarly, (Yuen *et al.*, 2016) proposed a deep neural network (DNN) for simultaneous facial landmark localization and head pose estimation. (Jha *et al.,* 2023) introduced a convolutional neural network (CNN)-based fusion framework that combines head pose and eye appearance to estimate the driver's gaze over a two-dimensional gaze grid. More recently, (Herashchenko & Farkaš, 2023) employed the MetaHuman platform to generate a synthetic facial dataset containing gaze and head pose annotations, demonstrating improved accuracy in eye gaze direction and head rotation estimation. While these studies show that localizing facial landmarks and estimating head pose are effective components of driver monitoring systems, most existing approaches focus on either directly regressing head pose using deep learning models or incorporating head orientation as an auxiliary feature in gaze estimation frameworks. Relatively little attention has been devoted, however, to designing geometric descriptors that exploit facial landmark configurations to characterize head pitch, particularly under frontal and lateral facial views, where self-occlusion limits the availability of facial landmarks.

To address these limitations, this paper proposes a multitask deep learning framework for thermal facial images that can perform facial landmark localization and head orientation classification (frontal or lateral view) simultaneously. The outputs of these two tasks are then combined to compute two sets of geometry-based descriptors for estimating head pitch, which are specifically designed for frontal and lateral facial views. The proposed framework adapts the pitch estimation process to the availability of visible facial landmarks under self-occlusion by selecting the appropriate descriptor according to the estimated facial orientation. These descriptors provide complementary geometric information that can be incorporated into thermal-based driver monitoring systems as an additional feature.

Metodology

2.1 Proposed Methods and Data Sources

This study presents a methodological framework based on a multitask CNN that simultaneously localizes facial landmarks and classifies head orientation. The objective is to estimate head pitch using geometry-based descriptors. Figure 1 illustrates the overall workflow of the proposed framework, which comprises four main stages: dataset partitioning, image and annotation processing, multitask network training and evaluation, and geometric head pitch estimation based on predicted facial landmarks and orientation.

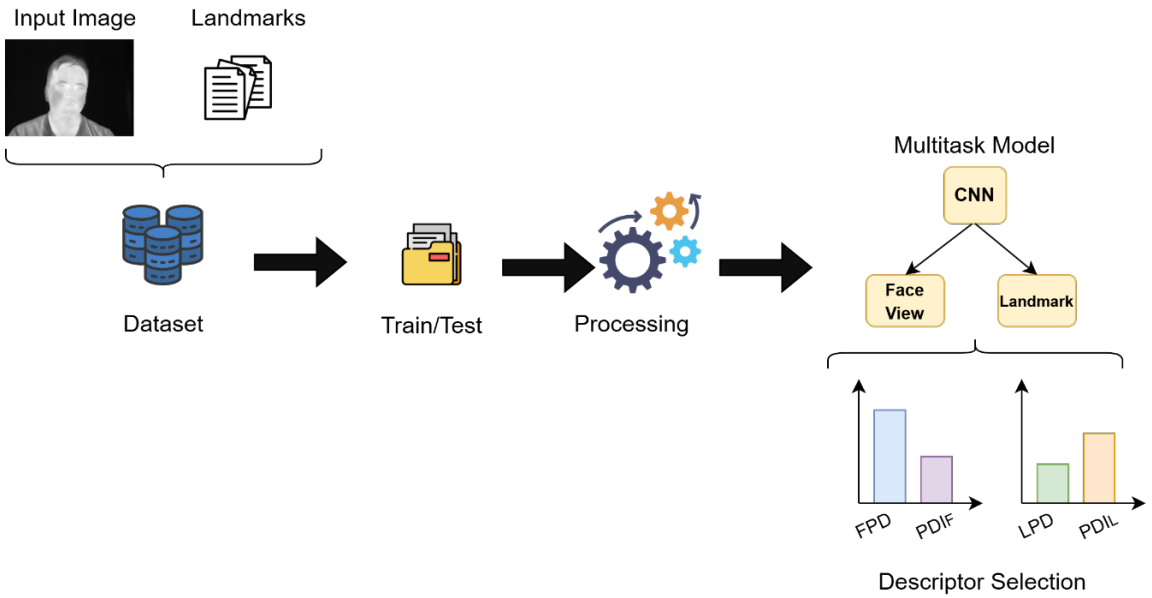


Figure 1. Framework of the studySource: Own work.

The proposed framework was evaluated using the Charlotte-ThermalFace dataset (Ashrafi *et al.,* 2022), which was collected by researchers at the University of North Carolina at Charlotte (UNC Charlotte). The dataset contains approximately 10,000 thermal infrared facial images acquired from ten participants under various environmental conditions, camera-to-subject distances, and natural head orientations. Along with the thermal images, the dataset provides environmental metadata collected during image capture, such as air temperature, relative humidity, airflow velocity, and camera distance. For this study, only manually annotated facial landmarks were considered. There are 73 landmarks for frontal views and 43 for lateral views. While the Charlotte-ThermalFace Dataset was not originally developed for head pose estimation, it is well-suited to the proposed methodology due to its substantial natural variations in facial orientation. These variations include frontal and lateral views, as well as different head pitch configurations. These characteristics enable the evaluation of the proposed geometric descriptors under realistic acquisition conditions without requiring dedicated head pose annotations. The complete annotated dataset is publicly available through the official repository: https://github.com/TeCSAR-UNCC/UNCC-ThermalFace.

2.2 Data splitting

The Charlotte-ThermalFace Dataset provides participant metadata, the (x, y) coordinates of each manually annotated facial landmark, and the image identifier that links each annotation to its corresponding thermal image. To evaluate the generalization capability of the proposed model, the dataset was partitioned at the subject level, using seven participants for training, one for validation, and two for testing. Unlike image-level random splitting, subject-wise partitioning prevents images from the same individual from appearing in multiple subsets, thereby reducing identity-related bias and providing a more realistic evaluation of the model's ability to generalize to previously unseen subjects.

2.3 Image Processing

2.3.1 Facial geometric analysis

Since the proposed geometric descriptors depend on spatial relationships among facial landmarks, JPEG thermal images from the Charlotte-ThermalFace dataset were used for facial visualization and landmark localization. These images preserve the thermal facial appearance while enabling the manually annotated landmarks to be accurately projected onto the facial geometry. Each image contains manual annotations corresponding to 73 landmarks in frontal views and 43 in lateral views, as illustrated in Figure 2. The reduced number of landmarks in lateral views is mainly due to self-occlusion, wherein several facial features become invisible due to head rotation.



a) b)

Figure 2. Manual annotations: a) 73 landmarks (front view), b) 43 landmarks (lateral view). Source: Adapted from Charlotte-ThermalFace (Ashrafi et al., 2022).

To standardize the input dimensions and reduce the computational complexity of the proposed deep learning model, the spatial resolution of all thermal images was resized to (256×256) pixels. The coordinates of the facial landmarks were proportionally rescaled to maintain their spatial correspondence within the resized images, as defined in (1).

$$x_i' = \frac{x_i * W_n}{W_o}, y_i' = \frac{y_i * H_n}{H_o} \quad (1)$$

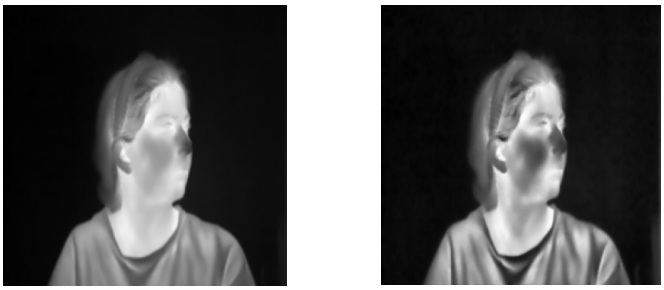
Where x_i and y_i denote the original coordinates of the i-th facial landmark, x_i' and y_i' are the corresponding rescaled coordinates, W_o and H_o represent the width and height of the original image, and W_n and H_n correspond to the dimensions of the resized image.

In addition to the landmark coordinates, head orientation labels were automatically generated for the classification branch of the multitask network. Since the Charlotte-ThermalFace Dataset provides 73 manually annotated landmarks for frontal facial views and 43 landmarks for lateral views, the number of available annotations was used as an indicator of facial orientation. Consequently, images containing 73 landmarks were assigned to the frontal-view class (Class 0), whereas images containing 43 landmarks were assigned to the lateral-view class (Class 1).

2.3.2 Contrast Limited Adaptive Histogram (CLAHE)

Contrast Limited Adaptative Histogram Equalization (CLAHE) is a local contrast enhancement technique commonly used in image processing to improve the visibility of the fine image details. Unlike conventional histogram equalization, which applies a global intensity transformation, CLAHE operates on small non-overlapping image regions (tiles) and independently equalizes the histogram within each region while limiting the local contrast amplification through a clip limit. This strategy reduces noise amplification while enhancing subtle intensity variations (Li *et al.,* 2024).

In the proposed framework, CLAHE was employed as a preprocessing step to enhance the local thermal contrast of facial regions before landmark localization. Improving the local contrast facilitates the convolutional network's extraction of discriminative facial features, particularly in areas with low thermal gradients. Based on preliminary experiments, the CLAHE parameters were set to clipLimit = 2.0 and tileGridSize = (8, 8). Figure 3 shows an example of the contrast enhancement obtained with the proposed preprocessing stage.



a) b)

Figure 3. Image processing: a) original image, b) image processed using CLAHE. Source: Adapted from Charlotte-ThermalFace (Ashrafi et al., 2022).

2.4 Model Construction and Evaluation

2.4.1 Multitask Network Architecture

Multitask learning (MTL) is a deep learning paradigm in which a shared feature extractor is optimized simultaneously to solve multiple related tasks via independent, task-specific output heads. Multitask models learn common feature representations, exploiting the correlations between related objectives while reducing the computational cost associated with training independent networks (Mao *et al.*, 2020). In the proposed framework, two related tasks are optimized simultaneously: facial landmark localization (the primary task) and facial view classification (the auxiliary task). Instead of directly improving landmark localization accuracy, the auxiliary branch provides the facial orientation information necessary for automatically selecting the appropriate geometric descriptor for subsequent head pitch estimation. Consequently, both tasks are learned together within a unified architecture, which enables end-to-end prediction of the information required by the proposed geometric framework.

The complete network architecture is summarized in Table 1. As shown, the network consists of a shared convolutional backbone made up of four convolutional blocks with 3x3 kernels. Each convolutional layer is followed by a max-pooling operation. The extracted feature maps are then flattened and projected into a fully connected layer with 512 neurons to generate a compact feature representation that is shared by both tasks. This shared representation is then divided into two task-specific branches. The landmark localization branch consists of a 10% dropout layer, followed by a fully connected layer with 146 output neurons corresponding to the normalized coordinates of the 73 facial landmarks. The facial view classification branch consists of a single

fully connected neuron with a sigmoid activation function that produces the probability that the input image belongs to either the frontal or lateral facial view.

***Table 1**. Structure of the Multitask Model.*

Layer (type)	Output Shape	Activation	#Params	Connected to
Input layer	**(256, 256, 1)**	-	**0**	-
Conv2D	**(254, 254, 16)**	**ReLU**	**160**	**Input Layer**
MaxPooling 2D	**(127,127, 16)**		**0**	**Conv2D**
Conv 2D_1	**(125,125, 32)**	**ReLu**	**4640**	**MaxPooling 2D**
MaxPooling 2D_1	**(62, 62, 32)**	-	**0**	**Conv2D_1**
Conv2D_2	**(60, 60, 64)**	**ReLU**	**18496**	**MaxPooling2D_1**
MaxPooling 2D_2	**(30,30, 64)**		**0**	**Conv2D_2**
Conv 2D_3	**(28, 28, 128)**	**ReLu**	**73,586**	**MaxPooling 2D_2**
MaxPooling 2D_3	**(14, 14, 128)**	-	**0**	**Conv2D_3**
flatten	**(25088)**	-	**0**	**MaxPooling2D_3**
Dense	**(512)**	**ReLU**	**12,845,568**	**flatten**
Head 1				
Dropout	**(512)**	-	**0**	**Dense**
Landmarks	**(146)**	**Sigmoid**	**74,898**	**Dropout**
Head 2				
Direction	**(1)**	**Sigmoid**	**25,089**	**flatten**

Since the proposed framework addresses both a classification task (facial view classification) and regression task (facial landmark localization) simultaneously, each output branch is optimized using a task-specific loss function within a weighted multitask objective. The landmark localization branch was assigned a weight of 0.8, whereas the facial view classification branch was assigned a weight of 0.2, prioritizing the primary regression task during joint optimization.

Facial view classification was formulated as a binary classification problem, where Class 0 corresponds to frontal facial views and Class 1 to lateral views. This branch was optimized using the Binary Cross-Entropy (BCE) loss function, defined as

Where y and $\hat{y}$ denote the ground-truth and predicted class label, respectively.

In contrast, facial landmark localization was formulated as a regression problem and optimized using a masked version of the Huber loss (Akhtar *et al.,* 2024) (Masked Huber Loss). Since lateral facial images only contain

43 visible landmarks, the remaining landmarks (44–73) are naturally absent due to self-occlusion. Consequently, the corresponding annotations in the original dataset are stored as NaN values.

Rather than removing these samples or forcing artificial target values, all missing landmark coordinates were temporarily replaced with −1 and subsequently ignored through a binary visibility mask during loss computation. This strategy prevents the network from being penalized for predicting coordinates associated with non-visible facial landmarks, allowing the optimization process to focus exclusively on valid landmark annotations. As a result, the regression branch learns only from visible facial structures while naturally handling landmark self-occlusion.

Let y_i and $\hat{y}_i$ denote the ground-thruth and predicted coordinates of the i-th landmark, respectively, and ley $m_i \in \{0,1\}$ represent the corresponding visibility mask, where 1 indicates a visible landmark and 0 denotes an occluded landmark. The visibility mask is defined as

$$m_i = \begin{cases} 1, & si\ y_i \neq -1 \\ 0, & si\ y_i = -1 \end{cases} \quad (3)$$

The absolute prediction error is computed as

$$e_i = |\hat{y}_i - y_i| \ (4)$$

The Hubber loss is then defined as

$$L_H(e_i) = \begin{cases} \frac{1}{2} e_i^2, & si\ e_i \leq \delta, \\ \delta\left(e_i - \frac{\delta}{2}\right) & si\ e_i > \delta. \end{cases} \quad (5)$$

Where $\delta = 0.05$ is the transition threshold between the quadratic and linear penalty regions.

Finally, the Masked Huber loss is obtained by considering only the visible landmarks and normalizing the accumulated loss by the number of valid landmark coordinates, To prevent numerical instability when no valid landmark annotations are available, the denominator is replaced by

$$\max\left(\sum_{i=1}^{N} m_i, 1\right) \quad (7)$$

2.4.2 Evaluation metrics

Since the proposed multitask network addresses both facial landmark localization and facial view classification, different evaluation metrics were used to assess each task´s performance. The landmark localization task was evaluated using Mean Absolute Error (MAE), which measures the average absolute difference between predicted and ground-truth landmark coordinates (Foster, 2022). The MAE is defined as

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (8)$$

Where N denotes the total number of evaluated landmark coordinates, y_i is the ground truth value and, $\hat{y}_i$ is the corresponding prediction.

The performance of the facial view classification branch was analyzed using a confusion matrix (Reschka *et al.,* 2022), which summarizes the classification outcomes in terms of correctly and incorrectly predicted samples. For the binary classification problem considered in this work, Class 0 corresponds to frontal facial views, whereas Class 1 represents lateral facial views, as illustrated in Figure 4.

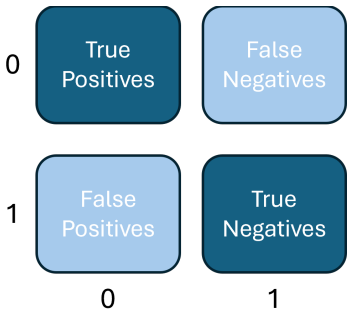


Figure 4. Confusion matrix structure.
Source: Own work.

From the confusion matrix, the overall classification accuracy was computed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

Where TP, TN, FP and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

2.5 Geometric-based head pitch estimation

The multitask network predicts facial landmarks that provide the geometric information necessary to estimate head pitch from thermal facial images. Since frontal and lateral views of the face exhibit different landmark configurations due to self-occlusion, two complementary geometric formulations were developed according to the estimated facial orientation, as illustrated in Fig. 5.

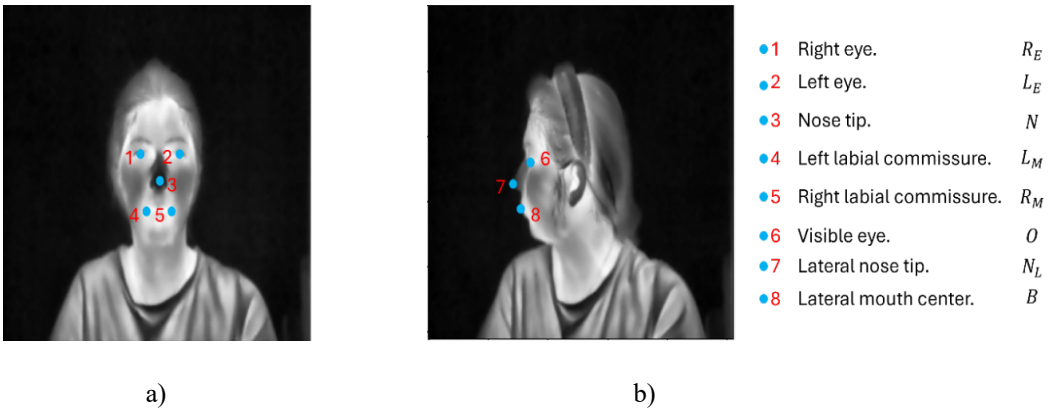


Figure 5. Keypoints used to estimate descriptors. a) Frontal keypoints, b) Lateral keypoints.
Source: Adapted from Charlotte-ThermalFace (Ashrafi et al., 2022).

Instead of directly regressing head pose angles, the proposed approach computes geometry-based pitch descriptors from the spatial relationships among facial landmarks. This strategy allows the estimation process to adapt to the subset of visible landmarks in each view while remaining robust to changes in facial scale and image resolution. All descriptors are formulated as ratios of Euclidean distances between anatomically meaningful landmarks to provide scale-invariant measurements of head pitch.

2.5.1 Frontal pitch Descriptor

For frontal facial views, the proposed descriptor is based on the observation that vertical head movements alter the distances between the eyes, nose, and mouth. As the head rotates upward or downward (pitch motion), the vertical facial proportions change while the anatomical relationships remain the same. Thus, the proposed descriptor estimates head pitch by analyzing the geometric configuration of these facial landmarks.

The frontal pitch descriptor (FPD) is computed using the landmarks illustrated in Figure 5(a). First, the centers of the eyes and mouth are obtained.

$$C_E = \frac{L_E + R_E}{2} \quad (10)$$

$$C_M = \frac{L_M + R_M}{2} \quad (11)$$

Where L_E and R_E denote the left and right eye coordinates, respectively, while L_M, R_M correspond to the left and right mouth corners. Consequently, C_E, C_M represent the centers of the eyes and mouth. The magnitude of the proposed frontal pitch descriptor is defined as the ratio between the distance from eye center to the nose and the distance from the nose to the mouth center,

$$d_{EN} = \left||C_E - N|\right| \quad (12)$$

$$d_{NM} = \left||N - C_M|\right| \quad (13)$$

$$FPD = \frac{d_{EN}}{d_{NM}} \quad (14)$$

Using a distance ratio instead of absolute distances makes the descriptor independent of facial scale and image resolution. A normalized Pitch Direction Indicator (PDI) is introduced to determine the direction of the pitch motion. First, the distance between the centers of the eyes and mouth is computed as

$$d_{EM} = \left||C_E - C_M|\right| \quad (15)$$

The normalized direction indicator is then defined as

$$PDI_F = \frac{y_N - \frac{yC_E + yC_M}{2}}{d_{EM}} \quad (16)$$

Where y_N, yC_E, and yC_M denote the vertical coordinate of the nose, eye center, and mouth center, respectively. The numerator resents the vertical displacement of the nose with respect to the midpoint of the eye-mouth axis, whereas normalization by d_{EM} makes the descriptor invariant to facial size. Negative values of the PDI_F indicate an upward head inclination, values close to zero correspond to a neutral pose, and positive values indicate a downward head inclination.

2.5.2 Lateral pitch Descriptor

In lateral facial views (Figure 5(b)), only one eye is typically visible due to facial self-occlusion. To overcome this limitation, a second geometric descriptor was formulated using only the landmarks that remain consistently visible in lateral views. The proposed Lateral Pitch Descriptor (LPD) is computed using the visible eye, nose, and center of the mouth. Similar to the frontal formulation, the LPD is based on the observation that the relative distances among these landmarks change when the head is tilted upward or downward. The corresponding Euclidean distances are defined as,

$$d_{ON} = \left||O - N|\right| \quad (17)$$

$$d_{NB} = \left||N - B|\right| \quad (18)$$

Where O denotes the coordinates of the visible eye, N corresponds to the nose tip, and B represents the center of the mouth. The magnitude of the LPD is then obtained as,

$$LPD = \frac{d_{ON}}{d_{NB}} \quad (19)$$

As in the frontal case, using a distance ratio makes the descriptor independent of facial scale and image resolution. To determine the direction of pitch motion in lateral views, the Pitch Direction Indicator (PDI) is defined as

$$d_{OB} = \left|\left|O - B\right|\right| \quad (20)$$

$$PDI_L = \frac{y_N - \frac{yO + yB}{2}}{d_{OB}} \quad (21)$$

Where y_N, yO, and yB denote the vertical coordinates of the nose, visible eye, and mouth center, respectively. Similar to the frontal descriptor, negative PDI_L values indicate an upward head inclination, values close to zero correspond to a neutral pose, and positive values indicate a downward head inclination.

Results

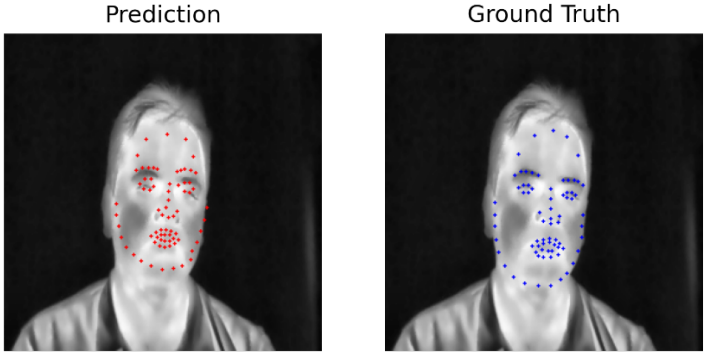
3.1 Landmark estimations

The performance of the facial landmark localization task was evaluated using the MAE, since this metric expresses the average localization error in the same units as the input image (pixels), providing an intuitive measure of the discrepancy between the predicted landmark positions and the corresponding ground-truth annotations. Table 2 summarizes the average MAE obtained on the training, validation, and test sets.

***Table 2**. Model performance in the landmark estimation task.*

Metric	Train	Validation	Test
MAE (pixels)	5.4739	7.6895	8.5052

As shown in the table, the proposed network achieved an average MAE of approximately 9 pixels on the test set. Although this result indicates the presence of localization errors relative to the manual annotations, the predicted landmarks preserve the overall facial geometry and the spatial relationships among key facial structures. Consequently, the estimated landmarks provide sufficient geometric information for computing the proposed head pitch descriptors. Representative examples comparing the ground-truth and predicted facial landmarks on the test set are presented in Figure 6.



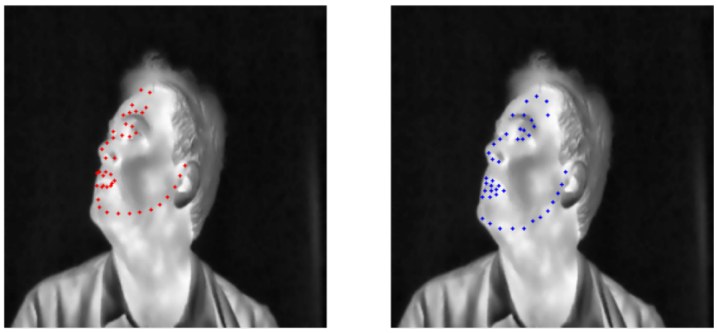


Figure 6. Comparison between estimated (left) and true (right) landmarks.
Source: Adapted from Charlotte-ThermalFace (Ashrafi et al., 2022).

3.2 Facial view classification

The performance of the facial view classification Branch was evaluated using a confusion matrix computed on the test set (Figure 7).

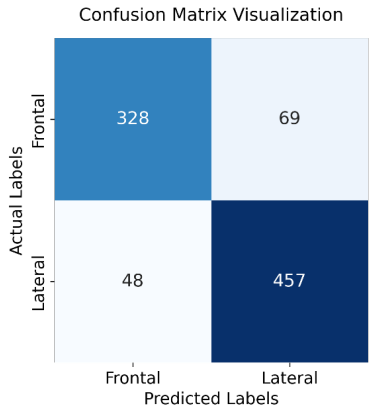


Figure 7. Confusion matrix results.
Source: Own work

The matrix shows that the multitask network successfully distinguishes between frontal and lateral views in most of the evaluated images, achieving 87% overall classification accuracy. This result shows that the shared feature representation can capture global geometric characteristics associated with facial orientation. It should be noted that the facial view labels were derived directly from the annotations provided in the Charlotte-ThermalFace Dataset. Images containing 73 annotated landmarks were assigned to the frontal-view class, whereas images with 43 annotated landmarks were labeled as lateral views, reflecting the landmark self-occlusion produced by head rotation. Most classification errors occurred in images exhibiting intermediate head rotations between frontal and lateral views. While these images retain the frontal-view label in the original dataset, the gradual loss of facial symmetry and partial self-occlusion of landmarks make them appear closer to lateral views. Consequently, the network tends to classify these ambiguous samples as lateral views. This suggests that some apparent misclassifications may actually arise from the dataset's discrete labeling strategy rather than from incorrect feature learning.

3.3 Head pitch descriptors results

The proposed head pitch descriptors were evaluated on the test set using the facial landmarks predicted by the multitask network, along with the estimated facial view. According to the predicted facial orientation, the corresponding geometric formulation (FPD/PFI_F or LPD/PDI_L) was applied to each image. Figure 8 shows representative descriptor values obtained under different head pitch conditions.

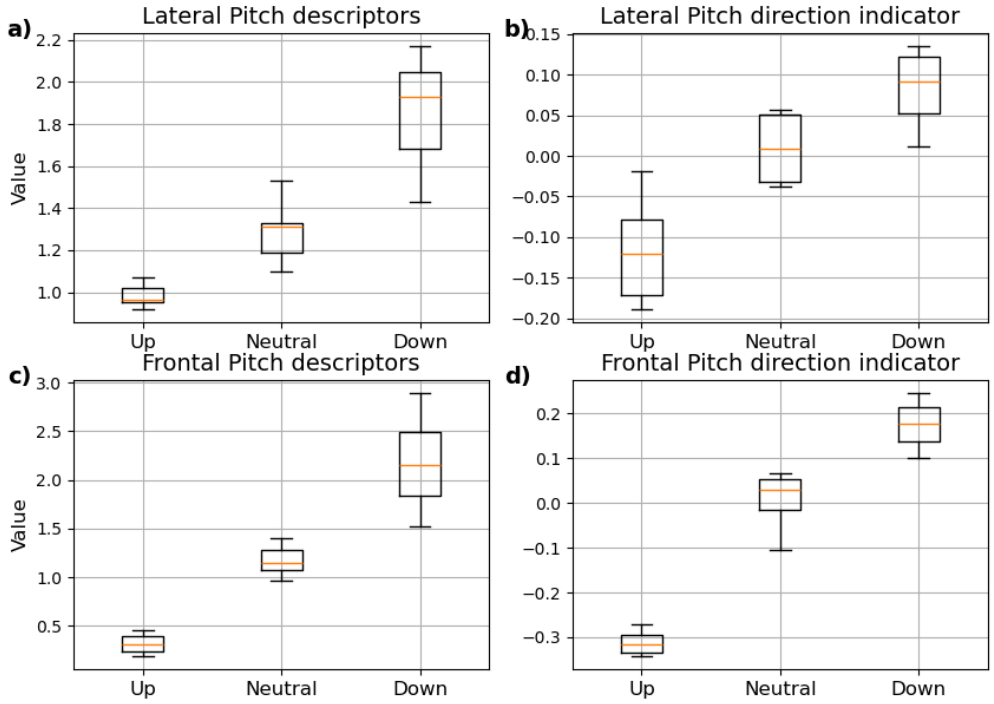


Figure 8. Behavior of pitch-related descriptors. a) Lateral descriptor (LPD), b) lateral direction indicator (PDI_L), c) Frontal descriptor (FPD), d) frontal direction indicator (PDI_F).
Source: Own work.

As shown in Figure 8, both FPD and LPD exhibit an increasing trend as the head rotates downward. Conversely, lower descriptor values are associated with upward head inclinations, while intermediate values correspond to near-neutral head poses. This behavior indicates that both descriptors successfully capture the geometric changes produced by pitch motion despite relying on different subsets of facial landmarks. Similarly, the proposed Pitch Direction Indicators (PDI_F and PDI_L) offer additional information about the direction of head movement. Negative values were consistently associated with upward head inclinations; values close to zero corresponded to neutral poses; and positive values indicated downward head rotations. Since both descriptors are normalized by facial distances, they are independent of facial scale and image resolution. This allows for consistent interpretation across different subjects.

Conclusions

This study presents a geometry-based framework for estimating head pitch from thermal facial images. This framework uses landmarks predicted by a multitask convolutional neural network. Unlike conventional approaches, which directly regress head pose angles, the proposed methodology estimates pitch using two geometric descriptors designed specifically for frontal and lateral facial views. This strategy allows the pitch estimation process to adapt to visible facial landmarks and naturally handle landmark self-occlusion. The multitask architecture efficiently predicts facial landmarks and the required facial view to automatically select the appropriate geometric descriptor. Although facial view classification is an auxiliary task, it allows both outputs to be integrated into a unified pipeline for estimating head pitch without requiring separate models or explicit landmark completion. The experimental results demonstrate that the proposed descriptors consistently behave the same way under different head-tilt conditions. This behavior provides complementary information that could be incorporated into driver monitoring systems as an additional feature for assessing fatigue or inattention. Despite the results, the accuracy of the proposed descriptors depends directly on the quality of the landmark localization and facial view classification. Consequently, localization errors or incorrect facial view predictions may propagate to the descriptor computation. Additionally, the evaluation was conducted

using a single thermal facial dataset with a limited number of participants, which may restrict the generalizability of the proposed methodology.

Future work will focus on improving facial landmark localization with more specialized deep learning architectures and incorporating normalized evaluation metrics, such as Normalized Mean Error (NME). This work will also involve validating the proposed descriptors on additional thermal facial datasets and integrating them into complete driver fatigue monitoring systems.

References

Abdic, I., Fridman, L., McDuff, D., Marchi, E., Reimer, B., & Schuller, B. (2016). Driver frustration detection from audio and video in the wild. *Proceedings of the KI,* 237. https://lexfridman.com/files/papers/2016_ijcai_abdic_fridman_driver_frustration.pdf

Akhtar, M., Tanveer, M., & Arshad, M. (2024). HawkEye: Advancing robust regression with bounded, smooth, and insensitive loss function. *arXiv*. arXiv:2401.16785.

Ashrafi, R., Azarbayjani, M., & Tabkhi, H. (2022). Charlotte-thermalface: A fully annotated thermal infrared face dataset with various environmental conditions and distances. I*nfrared Physics & Technology,* 124, 104209. https://doi.org/10.1016/j.infrared.2022.104209.

Caminos y Puentes Federales (CAPUFE). (2022, 21 de septiembre). Juventudes y accidentes viales. Gobierno de México. https://www.gob.mx/capufe/articulos/juventudes-y-accidentes-viales.

Dingus, T. A., Klauer, S. G., Neale, V. L. (. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R. (. R., Jermeland, J., & Knipling, R. R. (. R. (2006). The 100-Car Naturalistic Driving Study, Phase II - Results of the 100-Car Field Experiment (Report No. DOT-HS-810-593). *United States. Department of Transportation.* National Highway Traffic Safety Administration. https://rosap.ntl.bts.gov/view/dot/37370

Dong, Y., Hu, Z., Uchimura, K., & Murayama, N. (2010). Driver inattention monitoring system for intelligent vehicles: A review. IEEE transactions on intelligent transportation systems, 12(2), 596-614. doi: 10.1109/TITS.2010.2092770.

Foster, D. (2022). *Generative deep learning.* O'Reilly Media, Inc.

Herashchenko, D., & Farkaš, I. (2023, December). Appearance-based gaze estimation enhanced with synthetic images using deep neural networks. In 2023 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 129-134). IEEE. https://doi.org/10.48550/arXiv.2311.14175

Jha, S., Al-Dhahir, N., & Busso, C. (2023). Driver visual attention estimation using head pose and eye appearance information. *IEEE Open Journal of Intelligent Transportation Systems,* 4, 216-231. doi:10.1109/OJITS.2023.3258184

Kim, D., Park, H., Kim, T., Kim, W., & Paik, J. (2023). Real-time driver monitoring system with facial landmark-based eye closure detection and head pose recognition. *Scientific reports, 1*3(1), 18264. https://doi.org/10.1038/s41598-023-44955-1

Li, H., Wang, S., Li, S., Wang, H., Wen, S., & Li, F. (2024). Thermal infrared-image-enhancement algorithm based on multi-scale guided filtering. Fire, 7(6), 192.https://doi.org/10.3390/fire7060192

Li, N., & Busso, C. (2014). Predicting perceived visual and cognitive distractions of drivers with multimodal features. *IEEE Transactions on Intelligent Transportation Systems,* 16(1), 51-65. 10.1109/TITS.2014.2324414

Mao, L., Yan, Y., Xue, J. H., & Wang, H. (2020). Deep multi-task multi-label CNN for effective facial attribute classification. *IEEE Transactions on Affective Computing,* 13(2), 818-828. 10.1109/TAFFC.2020.2969189

Qiu, Y., Misu, T., & Busso, C. (2019, October). Driving anomaly detection with conditional generative adversarial network using physiological and can-bus data. In 2019 International Conference on Multimodal Interaction, 164-173.https://doi.org/10.1145/3340555.3353749

Raschka, S., Liu, Y. H., & Mirjalili, V. (2022). *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python.* Packt Publishing Ltd.

Yuen, K., Martin, S., & Trivedi, M. M. (2016, November). Looking at faces in a vehicle: A deep CNN based approach and evaluation. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) (pp. 649-654). IEEE.