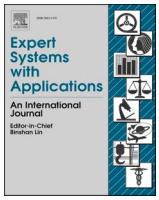
Expert Systems With Applications 254 (2024) 124425

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Deep face profiler (DeFaP): Towards explicit, non-restrained, non-invasive, facial and gaze comprehension

Wasiq Khan [a,*], Luke Topham [a], Hiba Alsmadi [b], Ala Al Kafri [b], Hoshang Kolivand [a]

[a] School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool, United Kingdom
[b] Department of Computing & Games, Teesside University, Teesside, United Kingdom

ARTICLE INFO

Keywords:
Comprehension tool
Facial analysis
Eye-tracking and gaze
Head pose estimation
Psychological profiling
Non-verbal behaviour

ABSTRACT

Eye tracking and head pose estimation (HPE) have previously lacked reliability, interpretability, and comprehensibility. For instance, many works rely on traditional computer vision methods, which may not perform well in dynamic and realistic environments. Recently, a widespread trend has emerged, leveraging deep learning for HPE specifically framed as a regression task; however, considering the real-time applications, the problem could be better formulated as classification (e.g., left, centre, right head pose and gaze) using a hybrid approach. For the first time, we present a complete facial profiling approach to extract micro and macro facial movement, gaze, and eye state features, which can be used for various applications related to comprehension analysis. The multi-model approach provides discrete human-understandable head pose estimations utilising deep transfer learning, a newly introduced method of head roll calculation, gaze estimation via iris detection, and eye state estimation (i.e., open or closed). Unlike existing works, this approach can automatically analyse the input image or video frame to produce human-understandable binary codes (e.g., eye open or close, looking left or right, etc.) for each facial component (*aka* face channels). The proposed approach is validated on multiple standard datasets, indicating outperformance compared to existing methods in several aspects, including reliability, generalisation, completeness, and interpretability. This work will significantly impact several diverse domains, including psychological and cognitive tasks with a broad scope of applications, such as in police interrogations and investigations, animal behaviour, and smart applications, including driver behaviour analysis, student attention measurement, and automated camera flashes.

1. Introduction

Facial profiling tools have increased in popularity in recent years due to their broader applications, including human–computer interaction (HCI) (Mukherjee and Robertson, 2015), psychological profiling such as deceptive behaviour (Khan et al., 2021), keepsakes (Yang et al., 2019), education profiling (Holmes et al., 2018), driver behaviour analysis (Venturelli et al., 2016; Mittal et al., 2016, 2016), nodding and shaking behaviour (Kong and Mbouna, 2015), multi-task learning (Huang et al., 2023), surveillance of crowd behaviour (Baxter et al., 2015), and many more (Murphy-Chutorian and Trivedi, 2009; Kumar et al., 2019). Generally, facial profiling tools (FPT) are based on eye gaze, head movements or poses, and other facial features (e.g., colour variation, blood flow etc.). Furthermore, gaze and head pose estimation (HPE) are interrelated topics and have been the main research focus of face profiling in various application domains (Liu et al., 2022; Liu et al., 2021). For example, (Liu et al., 2022) presents a three-branch HPE model using a matrix rotation to overcome estimation uncertainty. Similarly, (Liu et al., 2021) proposes Anisotropic Angle Distribution Learning (AADL) for HPE. The results demonstrated that AADL may overcome common issues such as motion blurring and incomplete data (Liu et al., 2021).

Growing interest in data-driven technologies, big data applications, autonomous systems, and smart applications needs more reliable and autonomous facial profiling. Furthermore, advancements in robotics, augmented reality, and smart city applications, such as driverless vehicles, increased the demand for comprehensive facial profiling tools in real-world scenarios. For example, robot-based assistance in various realistic scenarios has gradually emerged where head pose and eye gaze play vital roles in effective human-robot interaction and building trust

* Corresponding author.
E-mail addresses: w.khan@ljmu.ac.uk (W. Khan), l.k.topham@ljmu.ac.uk (L. Topham), H.Alsmadi@tees.ac.uk (H. Alsmadi), A.AlKafri@tees.ac.uk (A. Al Kafri), H.Kolivand@ljmu.ac.uk (H. Kolivand).

https://doi.org/10.1016/j.eswa.2024.124425
Received 5 December 2023; Received in revised form 3 May 2024; Accepted 4 June 2024
Available online 14 June 2024

with users during joint-attention tasks (Patacchiola and Cangelosi, 2017). Likewise, self-driving vehicles require reliable HPE and eye gaze estimation as fundamental components of driving assistance systems or driver's attention monitoring (Baxter et al., 2015).

HPE mainly comprises three-dimensional (3D) estimations (yaw, roll, pitch) and has been performed either by conventional computer vision (CV) algorithms or, recently, landmark estimation (Liu et al., 2022) and deep learning models (DL) (Liu et al., 2023). The traditional CV methods are based on composite geometric methods, appearance-based template matching, and features extracted from the face rectangle to categorise the head pose as one to three-dimensional problems (i.e., yaw, roll, pitch). Examples of such methods include geometric features-based key points detection (Diaz-Chito et al., 2016), morphable face model registration and optimisation algorithms (Meyer et al., 2015), 3D projection over 2D images (Kong and Mbouna, 2015), detector array using conventional machine learning (ML) (Moon and Miller, 2009; Rothwell et al., 2006), and feature descriptor (Ma et al., 2013). Alternatively, landmark estimation, such as DLib (King, 2009), OpenCV (Bradski, 2000), and recently (Liu et al., 2017), has been useful for conventional HPE and face detection methods. More specifically, these methods provide comprehensive information about micro-facial movements for close interactive applications such as psychological profiling. These methods utilise feature extraction, feature reduction, and ML methods to estimate the head pose in still images or video streams, which are affected mainly in outdoor environments. Furthermore, HPE from a single image frame is challenging and usually performed by mapping 2D images to 3D space, reducing the reliability as mapping the camera device properties is inconsistent for different videos.

Deep learning (DL) algorithms have recently outperformed the conventional ML approaches for facial analysis (Khan et al., 2021), facial recognition (Nanduri and Park, 2024; Himmi et al., 2024; Yin and Yu, 2024), face mask detection (Zeebaree and Kareem, 2023), and particularly HPE in real time for 2D and 3D images (depth as 3rd dimension). Various models have deployed DL for HPE for RGB images (2D), and RGB-D images (3D images) have been introduced. HPE has been modelled as a classification or regression problem for 3D HPE (rotation and movements) using convolutional neural network (CNN) and other variants of DL methods such as CNN trained over 3D images (Venturelli et al., 2016), multi-loss CNN using 3D image intensities (Ruiz et al., 2018), adaptive gradient CNN trained over detected faces (Patacchiola and Cangelosi, 2017; Yang et al., 2019) CNN trained over depth images (Mukherjee and Robertson, 2015), CNN based multi-model HPE (Hong et al., 2019; Ranjan et al., 2019) and adaptive gradient over LeNet-5 DL model (Patacchiola and Cangelosi, 2017). These methods indicate relatively higher accuracy, specifically when using depth images and ensemble models; however, utilising the depth images is not always feasible in real-world applications. A systematic review of existing HPE using conventional CV, landmarks, and DL methods can be found in (Kumar et al., 2019).

Similar to HPE, eye gaze estimation (EGE) and eye state estimation (ESE), such as blink, open close, etc., have been performed with a variety of domains such as security applications (Khan et al., 2021; Khan et al., 2021) healthcare technologies (Jyotsna et al., 2022), driver attention (Clark et al., 2019), and many more (Khan and Lee, 2019). Head and face detection is usually considered a pre-requisite to EGE and ESE, where iris location, pupil locations, and eye centre are estimated once the face is localised within the image frame. The eye gaze is then measured using pupil estimations and geometric calculations such as Euler angles. Similar to HPE, EGE and ESE have been performed mainly using conventional image processing (CIP) methods (Zeebaree and Kareem, 2023) and, recently, DL approaches (Murthy and Biswas, 2022). The literature shows the reliability of DL and landmark-based EGE (Amer et al., 2021) compared to conventional approaches.

Given the excessive research on HPE, EGE, and ESE methods with demanding applications scope, the existing methods lack in various aspects. Firstly, no reliable approach produces *explicit* facial profiling (e.g., all-in-one rotations, movements, and state for eyes and head) for offline or online video streams or image data. Secondly, the existing methods utilise either CIP, landmarks, or DL methods individually to perform HPE, eye tracking, EGE, and ESE, which could perform better using *hybrid* approaches. Specifically, the DL models mainly use depth information requiring special cameras, which are not always feasible specifically for outdoor dynamics and the mass of existing visual data (e.g., street surveillance cameras, YouTube videos, public visual data, etc.) in 2D form. Thirdly, most of the existing EGE and HPE methods work as regression models; however, considering the real-time applications, the problem could be formulated as classification (e.g., left, centre, right head pose) using a hybrid approach where data (video, images) needs annotations for varying head poses and eye gaze (e.g., up, down, centre, left, right etc.). The annotated data will then be used to train the composite model with the ability to extract the facial profile in both offline and online video streams and isolated images. Furthermore, the annotated dataset comprising class-level categorisation of eye gaze is publicly unavailable, requiring manual image annotations to train the hybrid models.

Considering the limitations of existing works specifically in relation to close-range HCI, such as psychological profiling and comprehension, we propose an explicit approach (DeFaP: Deep Face Profiler) for detailed facial profiling using a composite of the most recent techniques adopted from the DL methods, CIP, and facial landmarks for better accuracy, and newly introduced algorithms. The proposed DeFap can address the aforementioned limitations with the ability to extract comprehensive facial movements with potential uses for applications requiring precise analysis of psychological and interactive behaviour. Our major contributions in this work include:

a. *DeFaP method with the ability to process 2D RGB images and* videos *(offline and online) to generate comprehensive face profiles (in the form of binary codes) for facial movements precisely and reliably.*
b. *A hybrid model for HPE using transfer DL, landmark identifications, and newly labelled multi-class dataset.*
c. *A newly introduced algorithm for head roll estimation using geometric transformation.*
d. *A new dataset comprising 30,000 annotated iris instances covering diversity in terms of subjectivity, poses, gaze direction, and environmental dynamics (annotations will be provided upon request).*
e. *Eye gaze measurement using DL-based iris tracking model from 2D images and* video *stream.*
f. *An ESE using transfer learning utilising diversity in training and evaluation.*

Our work uses RGB images taken from monocular cameras, which permits the greatest portability in real-world applications, given the proliferation of such cameras in mobile phones and laptops. We use DL over annotated Iris images, labelled eye states (i.e., open or closed), and categorical head poses. Furthermore, we implement unconstrained gaze estimation, i.e., gaze estimation from a monocular RGB camera without assumptions regarding the user, environment, or camera.

The remainder of the manuscript is organised as follows: Section II presents the related works, followed by the proposed material, methods, and experimental design in Section III. Section IV presents the results of eye-tracking, ESE, and HPE, along with validation over multiple datasets. Section V discusses the outcomes and benchmarking of our results; Section VI concludes the proposed study and presents future research directives.

2. Related works

Psychological profiling through non-verbal behaviours has been investigated for many years and has various applications. Several terminologies have been used, including micro-movements, micro-

expressions, gestures, facial movements, etc., that are mainly extracted through HPE, EGE, and ESE, which are the focus of the proposed study.

2.1. Head pose estimation

The multi-dimensional HPE (i.e., rotation, movement) has been performed using various approaches with different applications. For instance, for comprehension analysis, (Holmes et al., 2018) modelled HPE as a discrete classification problem for different categories (e.g., left, right, centre). A similar approach is used in (Khan et al., 2021) with enhanced face detection followed by the array of binary classifiers for head rotations and facial movements (e.g., face up, face left, etc.). A study in (Diaz-Chito et al., 2016) presented HPE as a discrete classification problem for driver attention analysis. The head poses are categorised into the left, centre, and right, where geometric features are measured from the facial key points and forwarded to a conventional ML classifier. This classifier only works for the yaw rotations. Likewise, an unsupervised regression model is presented in (Drouard et al., 2017), indicating reliable outcomes for 3D estimations; however, it requires cropped face rectangles. Work in (Meyer et al., 2015) presented HPE by registering morphable face models to depth information while utilising an optimisation algorithm for efficient 3D registration. However, face detection is required prior to HPE. Likewise, an assumption about the camera focus point being known is made, which limits it further. (Kong and Mbouna, 2015) A 3D face morphing is proposed in (Kong and Mbouna, 2015) with depth parameters to estimate the head pose in 2D images. The disparity between 2D space and the projected 3D feature vector is reduced by rotating the reference 3D face model by pose angles of the query image. However, this method requires the query subject's 3D reference model (3D face) for the optimal HPE. Study (Yang et al., 2019) used 2D facial landmarks for HPE within original and deep-fake images. The estimated head poses from both groups' whole faces or central face regions are used to measure the alignment error (original vs. deep fake), representing the difference between head pose projections on an image plane. The calculated error is then fed to an ML algorithm to classify the original and deep fake instances. Furthermore, a detailed survey on conventional HPE methods can be found in (Murphy-Chutorian and Trivedi, 2009).

Recently, several studies have addressed the reliability of HPE utilising the DL models. (Venturelli et al., 2016)A DL-based HPE is used in (Venturelli et al., 2016) for in-car automotive driver attention and fatigue analysis, presenting a real-time HPE (10 frames/sec). The author used CNN trained over cropped face images to predict head pose as 3D motion. A study (Ruiz et al., 2018) proposed a multi-loss CNN regression model for HPE as 3D motion without utilising landmarks. Likewise, (Patacchiola and Cangelosi, 2017) used adaptive gradient methods with CNN to estimate the head pose (rotation and movement), where OpenCV is used for facial detection. (Yang et al., 2019) Spatial grouping of pixel-level features (FSA_Net) is proposed in (Yang et al., 2019) to form the region-level features used by multiple DL models (SSRNet ensemble with feature aggregation model) for HPE from a single frame. (Mukherjee and Robertson, 2015)A multi-modal HPE using regression over the multi-class outcomes from CNN is proposed in (Mukherjee and Robertson, 2015), which are trained over RGB-D images comprising pre-processed head regions of images. The outcomes indicate the usefulness of the multi-modal approach for both close-up faces and outdoor surveillance and environment interaction applications.

Deepgaze is an open-source library presented by (Patacchiola and Cangelosi, 2017) which uses multiple CNNs with adaptive gradient models for HPE (pitch and yaw) from RGB images. OpenCV is utilised to pre-process the face detector. Due to fast face detection, the system can process 15fps without GPU. However, the system indicates unreliability in real-time scenarios, specifically when the face detector mismeasures the face centre. (Hong et al., 2019) introduced a multi-task manifold DL face pose estimation (M^2DL) using CNN regression over 2D images. The face areas are extracted manually from the images to be used for the model training. Their approach indicated comparatively better performance in terms of multi-tasking, where HPE outcomes are primarily based on yaw. A survey on DL-based HPE is presented in (Kumar et al., 2019), while its applications in driver interaction and drowsiness detection are addressed in (Mittal et al., 2016, 2016).

While the existing works indicate the usefulness of the proposed HPE methods within the corresponding applications, the reliability and preciseness of conventional CV and ML methods in real-time dynamics need significant improvements (Mukherjee and Robertson, 2015). On the other hand, DL methods overcome these limitations; however, 3D movements as regression could be better modelled using a hybrid approach as in the proposed work. Furthermore, the depth information is not always feasible, which is considered in most of the existing DL-based HPE. Likewise, several DL-based HPE models do not consider HPE as 3D predictions. Likewise, they are more generic and do not only focus on the close-up HCI utilising RGB images. Sometimes, reliability is an issue when using RGB images such as (Patacchiola and Cangelosi, 2017). Furthermore, hybrid methods can better modulate comprehensive profiling for close group HCI. Detector-based head pose classification may be non-trivial due to low-quality images or misclassification of different poses from the same subject compared to the same head pose of different subjects (Mukherjee and Robertson, 2015). In other words, subjectivity must be considered in detector-based HPE such as (Rothwell et al., 2006).

2.2. Eye gaze and eye state estimation

Similar to HPE, EGE and ESE are vital for facial profiling and have been used in a variety of applications, including healthcare (Khosravan et al., 2019; Yiu, 2019; Medeiros, 2022), biometrics (Zemblys et al., 2018), behaviour analysis (Khan et al., 2021; Hickson et al., 2019); attention monitoring (Jiang et al., 2018; Kellnhofer et al., 2019; Jordan et al., 2020), and many more. The EGE can be performed generally using model-based (i.e., utilising the geometric model of the eye) or appearance-based methods (i.e., direct use of eyes as input). For instance, (Yu and Odobez, 2020) CIP-based 3D gaze estimation using gaze direction and warping field regularisation. However, it requires an eye image as a pair from the same subject, which is a limitation of this approach. The authors in (Holmes et al., 2018) presented a comprehension tool to extract facial micro-features comprising head and eye movements. They used detectors (array of ANNs) to classify the head and eye gaze. However, the approach has several limitations. For example, conventional ML and hand-crafted features might not generalise well. Likewise, CIP methods suffer from speculations, lighting conditions, and other real-time dynamics (Santini et al., 2018). Furthermore, tuning a more significant number of detectors is a challenge, specifically in generalisation. Recently, (Khan et al., 2021) used the Haar cascade to localise the face and eye region, followed by the detectors. However, it works only for the well-lit frontal faces with a clear background. Recently, (Khan et al., 2020) introduced a hybrid approach with facial landmark detectors and template matching for pupil localisation. The algorithm utilises a 2D convolutional cascade within the detected eye regions to identify the segment with the best matching score to a static kernel. The identified segment represents the pupil location within the detected eye region. This method indicated comparatively better performance for close group HCI expressly; however, this approach might suffer from light reflections easily and is limited to frontal pose only. While the model-based methods utilising CIP-based EGE and pupil localisation produced satisfactory performance in several applications, reliability (e.g., low image quality, varying lightening conditions), dependency on external sources (e.g., corneal reflection relying on the external light source), and generalisation (e.g., data diversity aspect) are the major challenges that are needed to be addressed with better methodologies.

In comparison, DL-based eye tracking and EGE (Khosravan et al., 2019; Yiu, 2019; Jiang et al., 2018; Kellnhofer et al., 2019; Choi et al.,

2019; Rakhmatulin and Duchowski, 2020; Krafka, et al., 2016; Park et al., 2018; Liu et al., 2021; Zhang et al., 2019; Fischer et al., 2018), more recently, indicated reliability and generalisation. An unconstrained gaze estimation approach (Gazenet) is presented in (Zhang et al., 2019), utilising facial landmarks and CNN for eye tracking. The face detection and facial landmarks are used to locate the landmarks within the input image. A generic 3D face model estimates the poses within the detected faces. The space normalisation method is used for segmentation and warping head poses and eye images to normalise training space. The outcomes indicate that head pose is vital for unconstrained gaze estimation. Similarly, (Kellnhofer et al., 2019) presented unconstrained gaze estimation (Gaze360). The LSTM models are used for the time progression learning to estimate the gaze uncertainty directly. The model indicated promising gaze estimation and its applicability for customer attention estimation in a supermarket setting. However, these methods are generally useful for broader perspectives, compensating for a higher degree of freedom, especially for eye gaze estimation.

Alternatively, constrained gaze EGE are especially useful for close-up faces with varying applications. For instance, DeepVog (Yiu, 2019) proposed gaze estimation for video-oculography in clinical neurology and neuroscience using CNN. However, the assumption of a spherical eyeball may be invalid in illuminations and other cases. Likewise, training and validation over diverse data (in terms of gaze and other data properties) would be useful for generalisation. The research (Rakhmatulin and Duchowski, 2020) proposed eye tracking with DL (YOLOv3) to classify the images into left, right, and centre gaze. In comparison, the study considered varying lighting conditions and standard validation datasets that could be used to validate. Likewise, training images comprise limited diversity. (Krafka, et al., 2016) A CNN-based eye tracking is proposed in (Krafka, et al., 2016) with a substantial image dataset (2.5 M images from 1450 subjects). Faces and eyes are cropped manually to train the model, whereas fixed locations are used to train the model. Likewise, (Park et al., 2018) proposed an intermediate pictorial representation of the input eye image to simplify the 3D EGE. However, this method has several assumptions about the average human eyeball iris geometry (e.g., size, diameter, shape), which are not always true in actual cases.

While most gaze detection uses pupil localisation, iris segmentation and localisation have also been used. For instance, (Jayanthi et al., 2021) presented iris segmentation using CIP and DL. However, they relied on CIP to annotate the iris, which might be unreliable in real-time dynamics and misidentification of the iris segment. The dataset is not annotated and, therefore, might produce an unreliable evaluation of the model performance. Recently, (Severo, et al., 2018) presented iris detection using DL over multiple datasets. The study concludes that DL-based iris detection is superior to CIP when validated over numerous datasets. The authors further presented annotations for the iris that are further extended within the proposed work to enrich the diversity, specifically regarding gaze direction, head pose, varying backgrounds, and subjectivity. Furthermore, YOLO infers the iris detection in real-time compared to CIP methods (Redmon and Farhadi, 2018). While state-of-the-art methods mainly use pupil localisation for gaze estimation, it has several associated challenges such as a) small size intolerant to dynamic situations (e.g., occlusions, light reflections, background noise); b) pupil dilation that occurs due to several factors including light illumination, age, emotion perception, recognition memory (Siegle et al., 2003) etc., which might cause imprecise pupil localisation that is vital specifically, for application requiring precise measurements such as security domain (Khan et al., 2021), healthcare (Yiu, 2019) etc., with close-up faces. The custom-trained YOLO iris detector would enable reliable localisation while resolving the challenges of natural dynamics such as noise, eyelids, eyelashes, background diversity and reflections.

Similar to iris detection, ESE has been performed using CIP approaches (Park et al., 2018) and, recently, DL methods (Medeiros, 2022; Jordan et al., 2020; Hu, 2020; Sanyal and Chakrabarty, 2019; Fogelton and Benesova, 2018; Ryan, 2021). For instance, (Ibrahim, et al., 2021) used the Haar cascade to detect the face from the input image, followed by facial landmark detection to crop the eye region. The eye state is then estimated using the eye-aspect ratio. However, such an approach might suffer from dynamics and diverse conditions such as varying poses, gaze, occlusions, light reflection, and, specifically, varying thresholds for variable conditions, which is impractical in most applications. In contrast, DL-based ESE indicated better accuracy and reliability in handling such conditions. For instance, (Hu, 2020) used LSTM for the eyeblink in the wild, and (Jordan et al., 2020) used CNN to detect eyeblink for driver drowsiness. Moreover, (Sanyal and Chakrabarty, 2019) used CNN-based ESE utilising diverse datasets to train and validate their approach, (Medeiros, 2022) presented eye blink dataset and ML-based ESE for Amyotrophic Lateral Sclerosis patients where they used moving average for the blink detection, (Fogelton and Benesova, 2018) used RNN-based ESE and evaluated on diverse datasets, and (Ryan, 2021) proposed driver monitoring tool using ESE using YOLO regression model.

We present a complete comprehension method enabling both head and eye movements with 20 movements (facial codes) while considering the reliability and validation in real-time data extraction. We utilise hybrid methods comprising DL, automated landmarks extraction, and CIP to extract and validate the outcomes during the data generation. We further make the annotations available for the research community to be used for other related applications.

The remaining manuscript is organised as follows: Section III presents the proposed methods and material, and Section IV presents the results of both Eye-tracking and HPE. Section V discusses the outcomes, and Section VI concludes the proposed study.

3. Material and methods

The proposed DeFaP approach comprises three main components, including A) HPE model, B) iris detector, and C) ESE, which are then further embedded within the facial landmark extractor to encode a fully autonomous profile of facial movements from a real-time video stream or input image frame. Next, D) provides an overview of the proposed DeFaP method, which combines the components described in A), B), and C). This section describes the datasets and methods used in these components and the detailed implementation for each task.

3.1. Head Pose Detector (HPE)

HPE is performed using a detector approach while exploiting deep transfer learning (DTL) for the custom training of multiple DL models. Generally, DTL involves reusing a previously trained model applied to a new problem. There are a variety of possible methods, mainly including multi-class models and multiple binary models. We used the former for the sake of simplicity. This section will first introduce the constructed dataset used to train the models used in this work. Secondly, a description of a bi-model classification approach to discrete HPE using DTL. Thirdly, a simple mathematical approach to estimating head roll is introduced.

1) Dataset Preparation for HPE

The dataset used for HPE is constructed using three existing datasets: BIWI (Fanelli et al., 2013), UPNA (Ariz et al., 2016), and UPNA Synthetic (Larumbe et al., 2017). Each dataset provides images of a single face with the associated rotation angles. The combination of these datasets provides 35 participants, of which ten are synthetic, and approximately 42,000 images. Considering the data diversity and model generalisation, we use the first dataset (i.e., BIWI dataset) for training purposes while preserving 20 % of the BIWI participants for evaluation/validation purposes (i.e., the participants in this set are not present in the training set). We further evaluate the trained model over two additional

purely unseen datasets (i.e., UPNA and UPNA Synthetic) to validate the generalisation of the trained model.

To label the aforementioned datasets, each head pose instance is categorised into one of nine head pose categories (as shown in Fig. 1-a): Up-Left, Up-Centre, Up-Right, Centre-Left, Centre-Centre, Centre-Right, Down-Left, Down-Centre, and Down-Right. Each head pose instance is placed in the relevant category based on the rotation angles provided by the dataset, as per the boundaries provided in Table 1.

Next, the constructed dataset required pre-processing to ensure consistency. Due to the necessary combination of multiple datasets, instances were recorded at different distances from the camera. Therefore, extracting the faces from the background images was necessary. The faces were extracted from the categorised images using the face detection component of MediaPipe Face Mesh (Google, "MediaPipe Face Mesh," GitHub, 2020). The extracted faces were stored as $128 \times 128 \times 3$ PNG image files. A pre-processed example from each category is displayed in Fig. 1(a).

2) Bi-model classification

This subsection describes the implementation of a bi-model solution to HPE using DTL techniques. Training a DL model is expensive in terms of both time and computational resources, particularly in our case, which requires multiple models. Hence, there has been a rise in the popularity of DTL, a deep learning technique for transferring knowledge from one model to another to solve a related problem. It is an appropriate foundation for the HPE solution, considering the benefits of DTL. Firstly, the nine-category HPE problem is broken into two halves: vertical estimation, up, centre, and down, and horizontal estimate, left, centre, and right. The combination of these two halves provides all nine of the desired categories. We utilise VGG19 (pre-trained model) for this purpose while fine-tuning over custom data to model HPE as a detector array. However, as the constructed dataset is annotated with nine categories for the overall HPE problem, the dataset must first be modified before it can be used to train the two sub-problems.

To provide training data in a format relevant to the two-model approach (from the BIWI dataset), they were duplicated to provide a vertical and a horizontal training set. In the vertical training set, the original categories are merged into three: up (up-left, up-centre, up-right), centre (centre-left, centre-centre, centre-right), and down (down-left, down-centre, down-right). Similarly, in the horizontal training set, the original categories are merged into three: left (up-left, centre-left, down-left), centre (up-centre, centre-centre, down-centre), and right (up-right, centre-right, down-right). Fig. 2 visualises this arrangement; the rows represent the vertical groups, and the columns represent the horizontal groups. Fig. 3 shows the implementation of the proposed bi-model HPE.

Table 1
Boundaries for Head Pose Categories.

Category	Pitch Range (Deg)	Yaw Range (Deg)
Up-Left	>15	<-15
Up-Centre	>15	<=15 & >=−15
Up-Right	>15	>15
Centre-Left	<=15 & >=−15	<-15
Centre-Centre	<=15 & >=−15	<=15 & >=−15
Centre-Right	<=15 & >=−15	>15
Down-Left	<-15	<-15
Down-Centre	<-15	<=15 & >=−15
Down-Right	<-15	>15

3) Head roll estimation (HRE)

The proposed method of discrete HPE does not consider the roll rotation angle. A simple mathematical method of calculating head roll is proposed to compensate for this. In this method, two key points are identified in the image; the simplest key points are the tip of the nose and the centre of the forehead, as shown in Fig. 1 (B). Other key points can be used in this calculation; however, the calculation is relatively simple as the nose tip is near the face's centre, and the forehead's centre is usually vertically aligned with the nose when the head roll is zero. Other key points may be used when the desired key points are occluded. The calculation aims to calculate the angle between the two lines: *a*, a vertical line drawn from the nose key point, and *b*, the line between the nose and forehead key points. For convenience, lines *a* and *b* are vectorised as *u* and *v*. *Cos*Θ can be calculated using the cosine formula from Eq. (1) (Neill, 2018). Θ can then be extracted using the inverse cosine rule from eq (2) (Neill, 2018) to provide an estimate of the head roll rotation angle. The proposed HRE is evaluated using the BIWI dataset, which comprises the annotations for the current state of the head pose.

$$\cos\theta = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\cdot\|\vec{v}\|} \tag{1}$$

$$\theta = \cos^{-1}\left(\frac{a}{b}\right) \tag{2}$$

3.2. Iris detector

1) Iris dataset preparation

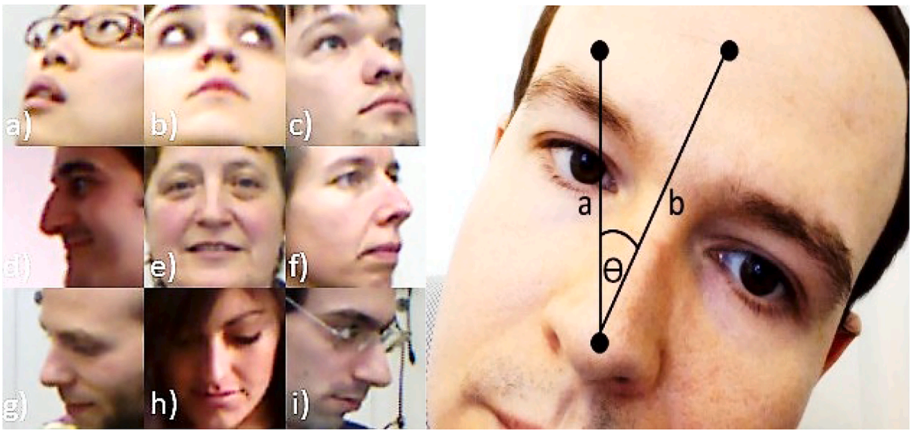


Fig. 1. An example pre-processed image from each head pose category: a) Up-Left, b) Up-Centre, c) Up-Right, d) Centre-Left, e) Centre-Centre, f) Centre-Right, g) Down-Left, h) Down-Centre, and i) Down-Right. B) Head roll angle Θ calculated as the angle of the line segment ab where a is a vertical line originating at the tip of the nose and b is the line between the nose's tip and the forehead's centre. The head pose category images (left side) are edited from the public dataset (Fanelli et al., 2013), and consent is provided by the head roll participant (right side).

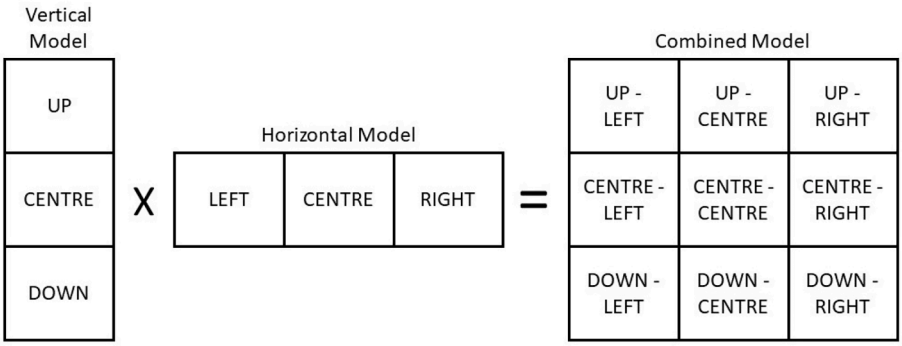


Fig. 2. Vertical and horizontal groups combined to produce their relevant sub-categories. For example, UP (from the vertical model) and LEFT (from the horizontal model) produce a combined UP-LEFT.

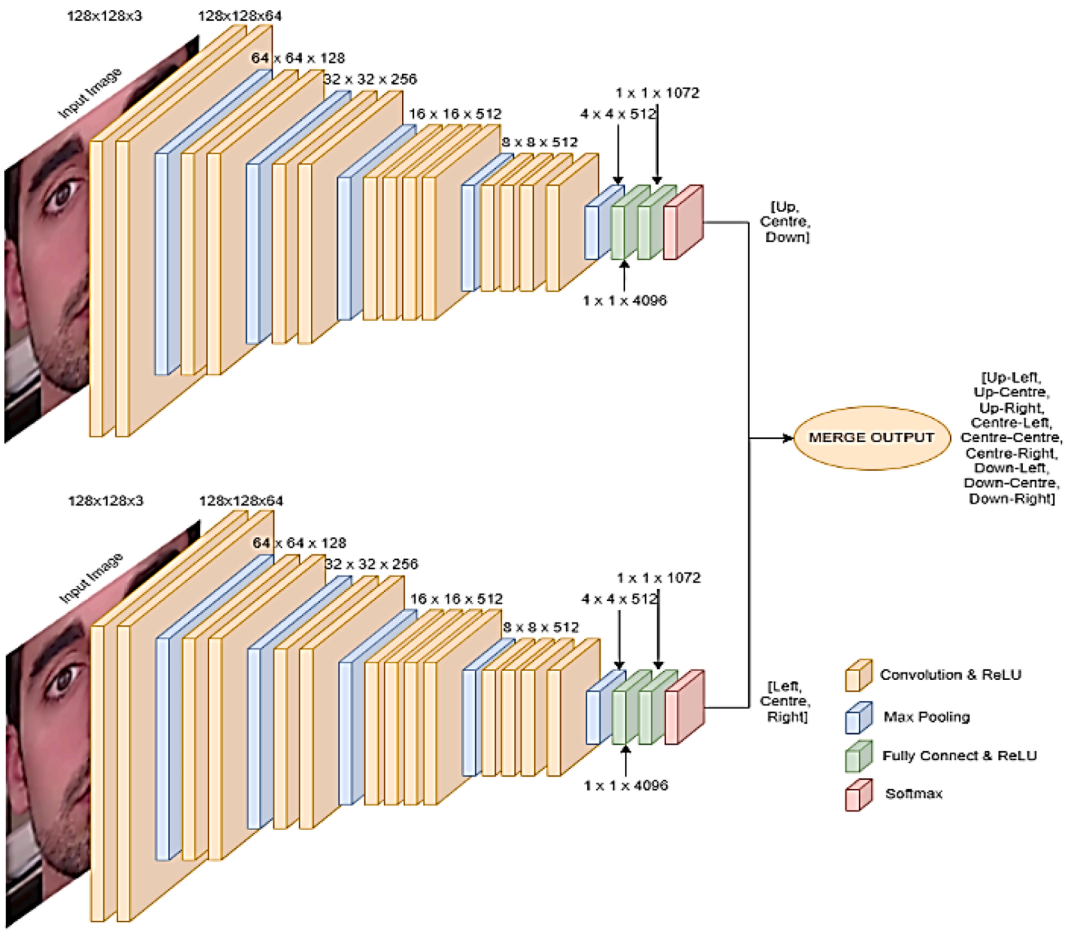


Fig. 3. Bi-model implementation of a VGG-19-based deep learning model for discrete HPE. The vertical model identifies whether the subject's face is pointing up, centre, or down, and the horizontal model identifies whether the subject's face is pointing left, centre, or right. The outputs are combined to produce nine categories.

We collected the dataset comprising diverse head movements and eye information (e.g., pupil, iris) from various public sources, including the Columbia gaze dataset (Smith et al., 2013), GazeCapture (Krafka, et al., 2016), GI4E database (Villanueva et al., 2013), and head pose database UPNA (Ariz et al., 2016). As mentioned earlier, the eye-tracking dataset comprises pupil information; however, for the localisation of DL-based eye-tracking, the available datasets are either not annotated or contain limited diversity. We prepared a combined dataset comprising iris annotations from the aforementioned datasets.

As shown in Fig. 4, the input images (or video frames) are processed using DLib (King, 2009) (public library) for the facial landmark extraction (i.e., face and eye region). The raw dataset is fed into landmark extraction to locate facial and eye segments and remove invalid image frames. For instance, a frame is not further processed (i.e., invalid frame) unless an appropriate count of face and eye segments exists. This filters out a lot of unnecessary background that may cause false positives for iris detection. The extracted eye segments are then processed further to perform the manual annotations.

In total, we annotated over 15,000 image frames (producing over 30,000 individual iris annotations) from 145 individuals. Concerning iris location, the annotated data comprises 2492, 9081, and 2492 annotated frames for right, centre and left iris positions (i.e., target classes). The newly prepared dataset ends up with both images and videos collected from a combination of high-quality lab conditions and mixed-

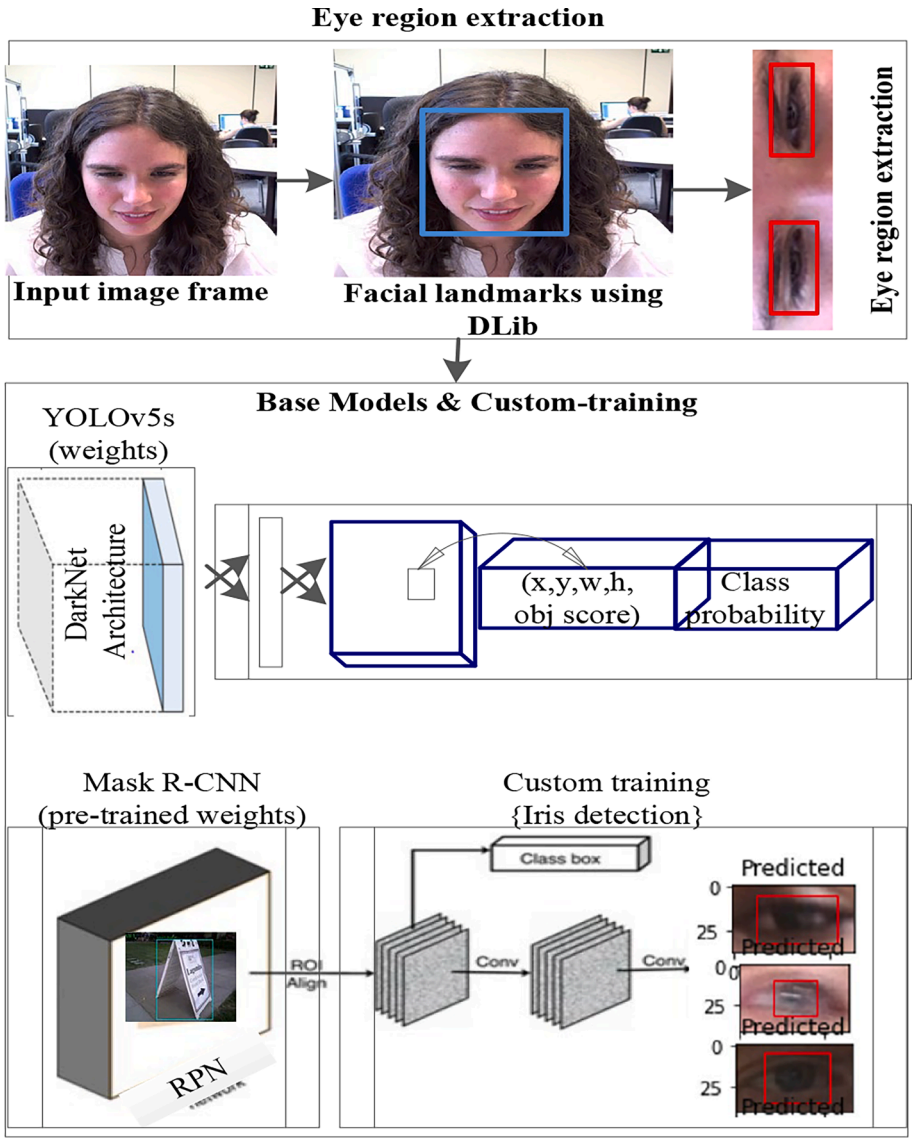


Fig. 4. Building blocks of the proposed iris detection approach. The top layer uses facial landmarks to extract the eye region and annotate the iris. At the same time, the second block performs the iris detection within the extracted eye segments using transfer learning.

quality crowd-sauced conditions. The annotated data comprises varying image sizes, resolution, subjects, perspectives, eye colour, eye features, glass-wearers, iris locations, etc., to make the dataset more diverse and valuable for generalising the proposed system. A detailed description of each dataset, acquisition settings and other information can be found in Supplementary S1 (Table S1).

2) Iris detection

As mentioned earlier and discussed in (Khan et al., 2023), DTL leverages pre-trained models to enhance processing efficiency and generalisation and eliminate traditional approaches' need for the initial training time. Several pre-trained models are available for object detection, including YOLO (Redmon and Farhadi, 2018), Faster-RCNN (Ren et al., 2017), and Mask R-CNN (He et al., 2017), which is an instance segmentation DL model designed to identify multiple objects within an image frame. In addition to the bounding boxes and class names, Mask R-CNN provides masks for the resulting image. Mask R-CNN's first component generates region proposals (RPN) for each object within the input image. In contrast, the second step generates class-level information, corresponding bounding boxes, and the pixel-level mask for identified objects based on RPN information. The Feature Pyramids Network (FPN) (Yang et al., 2017) forms the backbone of Mask R-CNN, used for object detection in images of varying scales. FPN's variable scale maintains robust semantic features compared to a single CNN. The original work presents further details on Mask R-CNN implementation, and mathematical formulation (He et al., 2017).

YOLO, which utilises a single CNN to detect object positions and corresponding classes, outperforms the R-CNN family with multi-stage processing (Plastiras et al., 2016), which means that YOLO performs object detection as a single regression task directly from the input image to the predicted objects' locations with associated class probabilities. YOLO has been updated several times (Redmon and Farhadi, 2018) with the latest version, YOLOv5, which was trained over the COCO dataset. It has three variants: small, medium, and large networks. Despite its fast performance, YOLO has some limitations. For example, its performance deteriorates when detecting small-sized objects or objects too close to each other in an image (Cao et al., 2021).

Considering the problem in hand (i.e., iris detection) and the aforementioned pre-trained deep models, we utilised mask R-CNN and YOLOv5s for the hyper-tuning over our custom dataset comprising iris annotations (Section III-B-1). As shown in Fig. 4, we used the pre-trained

weights for both models while training the fully connected layer over the custom dataset to utilise the prior knowledge of the existing model trained over larger datasets. The custom-trained iris detection models are available online and can be accessed upon request.

3.3. Eye state detector

1) Eye state dataset

The data set used to train the models in the first instance was the MRL dataset (Fusek, 2018) containing 4,545 images, 2355 closed-eye images, and 2190 open-eye images. The MRL image size was 24 x 24, pre-cropped around the eye, so the eye was centred within the image. The dataset was divided into two subsets: the training and the test datasets. The training dataset represents 80 % of the entire dataset and 20 % of the test dataset for the validation of the networks. We use the mEBAL dataset (Daza et al., 2020) comprising 4,252 images for further validation over a purely unseen dataset captured in different settings and using other protocols. This dataset contains 3256 open-eye images and 996 close-eye images. The format of the mEBAL images is 24 × 24 pixels, which were pre-cropped around the eye so that the eye is centred in the middle.

As the datasets are captured in realistic environments, they must filter out the noisy instances. However, both datasets contain diversity in various aspects. For example, variations in lighting conditions where darker images will prove more difficult for the models to identify. Likewise, there are various eye types (e.g., people with an epicanthic fold), glasses wearers (that may reflect light differently, which is more challenging) and background variations.

2) Eye state estimation

Similar to HPE, as described in Section III-A, we utilised DTL using VGG19 pre-trained weights to extract features from input eye images. The extracted features are fed into a fully connected layer to classify the input image frame into ‘open-eye’ and ‘closed-eye’ states. Detailed configuration of the custom-trained ESE, along with the trained ESE model, is available online and can be accessed upon request.

3.4. Proposed DeFaP method

Algorithm 1 describes the end-to-end procedure of extracting facial and gaze codes from the video frames using HPE, HRE, Iris detection, and ESE, which are the components of the proposed DeFaP method. The outcomes from DeFaP will be stored in a vector form comprising binary codes, representing the corresponding state (e.g., eye is open or closed; head pose is left, right, top-left, etc.) for each channel (i.e. head, eyes, face). In the first step, a pretrained library DLib (Meyer et al., 2015) is used to extract the facial landmarks from the input image frame, which are used to localise facial channels, including full face, eyes, nose etc. An input frame is considered valid if it contains a face. Otherwise, the next frame is processed. For each valid input frame, the head roll is estimated from the cropped head segment. Then, the custom-trained HPE model is utilised to identify the head state (e.g., left, right, down, etc.) within the current frame. We then automatically crop the eye regions using facial landmarks (from DLib) to confirm the presence of exact two eyes within each identified face segment. In the positive case (i.e., exactly two eyes are identified), the custom-trained ESE model is used to classify the current states of both eyes (as open or closed). If an eye state is found to be ‘open’, the custom-trained iris detection model is then utilised to estimate the iris centre, which is then used for the eye gaze estimation (e.g., looking left, up, right, etc.). This procedure recursively continues to process each frame of the input video, and the outcomes are stored in a linear vector representing the binary state (i.e., binary code) for each channel of information.

Algorithm 1. Face and gaze comprehension data generation from video **and image input using the proposed DeFaP**

Input: $V = \{v \mid v \text{ is a video}\}$ or an image frame f to be processed by DeFap
Output: *DeFaP vector* $\Rightarrow dP \in \{HP, EG, ES\}$*where* :
- $HP = \{Left, Right, Center, Top, Bottom, topLeft, topRight, bottomLeft, bottomRight\}$*istheheadpositionset*;
- *HRistheheadrollangle in a specific f*
- $EG = \{Left, Right, Center, Top, Bottom, topLeft, topRight, bottomLeft, bottomRight\}$*istheeyegazesetforspecificf*;
- $ES = \{open, close\}$*istheeyestatesetforspecificf*

Process:
Step 1 (head roll estimation)
For each f of input video v:
i. Extract facial landmarks from the input frame f using pre-trained methods such as DLib (Meyer et al., 2015) or OpenCV (Moon and Miller, 2009)
ii. Set a = forehead marker from f
iii. Set b = nosetip marker from f
iv. Measure the $HR = \cos^{-1}\left(\frac{a}{b}\right)$ for current frame f
Step 2 (head pose detection):
IF the current frame f contains a face (i.e., if there exists a face in step i):
v. Set validFace (a Boolean variable) as *true*
vi. Segment the face rectangle (F_R) using the face coordinates from step i
vii. Segment the head region H_r using F_R
viii. Extract the head pose using the proposed HPD (Section III-A) and store the outcomes in output vector *dP*
Step 3 (eye segmentation):
IF F_R contains two eyes
ix. Set validEyes as *true*
x. Segment the eye frames (E_L, E_R) for the left and right eye from the current f(using facial landmarks in Step 1-i)
Step 4 (eye state estimation):
For each eye segment E_L, E_R:
xi. Detect the eye state using the proposed ESE model (Section III-C) and custom gaze threshold and store the outcomes in *dP*
Step 5 (iris detection):
IF eye state is *Open:*
xii. Estimate the iris location using the proposed Iris detection model (Section III-B)
xiii. Measure the iris centre (I_c) using localised iris
xiv. Transform the I_c into eye gaze and store the outcomes in output vector *dP*
ELSE
xv. Transform the I_c into eye gaze and store the outcomes in outcome vector *dP*
End loop
ELSE
xvi. Set validEyes as false
ELSE
xvii. Set validFace as false
END loop

3.5. Experimental setup

Multiple experiments (Exp) are conducted considering the three major components of the proposed DeFaP and are detailed as follows.

- *Exp 1_a. HPE. The model is trained over 80 % of the BIWI dataset while testing over 20 % of the dataset while considering the subjectivity (Leave-K-Out strategy).*
- *Exp 1_b. HPE. The model is trained over the BIWI dataset and cross-validated over the UPNA and UPNA-Synth datasets.*
- *Exp 1_c. HPE. The proposed head roll is evaluated using the BIWI dataset.*
- *Exp 2_a. Iris Detection. Mask RCNN and YOLOv5 models are trained over 80 % of the annotated dataset and validated over the rest of 20 %.*
- *Exp 2_b. Iris detection. Mask RCNN and YOLOv5 models are trained over annotated datasets and tested over cross-datasets.*
- *Exp 3. ESE: The model is trained over 80 % of the combined training dataset (Section III) and evaluated over 20 % of the unseen instances, as well as additional dataset* (Jung et al., 2017).

Performance for the above experiments is evaluated using various gold standard metrics as appropriate in each case. For classification (e.g., HPE. ESE), we utilise standard metrics, including accuracy, sensitivity, and specificity, whereas we use MAE for the regression task (i.e., HRE). Likewise, we employ mean average precision (mAP), recall, and

precision for iris detection, which are the standard metrics for object detection. Within the context of this study (and corresponding task), these metrics are defined as follows:

TP: Correctly classified images that belong to that class; **TN**: Correct rejection of images that do not belong to that class; **FP**: Incorrect classification of images to a class they do not belong to; **FN**: Incorrect rejection of images from a class they belong to.

Recall or Sensitivity $_{(c)}$: $\frac{TP_C}{TP_C+FN_C}$, percentage of images that were classified to class C, compared to all images that should have been classified into C.

Specificity$_{(c)}$: $\frac{TN_C}{TN_C+FP_C}$, percentage of negative instances or true negatives (TN) out of all actual negative instances.

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$, overall accuracy of the model for all classes.

Precision$_{(c)}$: $\frac{TP_C}{TP_C+FP_C}$, percentage of images correctly classified for class *C*.

Macro Average (F1 Score): $\frac{\sum_{c=1}^{n} F1\ Score_{(C)}}{n}$, average of each class's F1 score independent to sample size per class where, F1 score is Harmonic-mean of precision and recall indicating success rate of the model for class C.

mAP (mean average precision): $\frac{1}{N}\sum_{i=1}^{N} AP_i$, average Precision (AP) for each class (i.e., 1 to N) and then average over the number of classes (N).

MAE (Mean Absolute Error): $\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\widehat{y}_i\right|$, average difference between the actual values ($y_{i..n}$) and the predicted values ($\widehat{y}_{i..n}$) in a regression task.

Further details and mathematical formulations of these metrics are presented in (Ren et al., 2017). For the experiment (data preparation, DL models' training and validation), we use a GPU machine comprising 4× AMD Ryzen Threadripper 2990WX, 2 TB hard drive, 128 GB RAM, and 32-core 3.00 GHz Intel CPU.

4. Results

This section describes the HPE, HRE, EGE and ESE results using the datasets and experimental setup detailed in Section III. We obtained statistical results for custom-trained DTL-based classification and detection models. To ensure standard performance metrics were measured, we followed the optimal configurations recommended by the original sources (Redmon and Farhadi, 2018; He et al., 2017).

4.1. Discrete head pose estimation (Exp 1_a, Exp 1_b)

As mentioned previously, 20 % of the BIWI dataset (Fanelli et al., 2013) participants were withheld from the training dataset to evaluate the HPE method. As shown in Table 2, the accuracy when estimating all nine categories using the BIWI evaluation set is 86.79 %. This is lower than the 99.21 % accuracy reported in (Elharrouss et al., 2020); however, this solution estimates only three categories (left, centre, and right) compared to nine in the proposed HPE approach. Also shown in Table 2 are the accuracies of the individual models; the horizontal model achieved 92.75 %. At the same time, the vertical model achieved 89.04 % accuracy. As a correct estimation requires both models to provide accurate overall estimates, the maximum accuracy is limited to that of the lowest-performing model.

The same VGG-19 approach was adapted to estimate all nine categories using a single model to compare the two-model approach. When evaluated using the same BIWI evaluation set, the single-model system achieved just 70.1 % accuracy, significantly lower than the 86.79 % achieved when combining two models, suggesting that the two-model approach provides considerably higher accuracy.

Furthermore, Table 3 presents the results of the two-model approach evaluated using previously unseen datasets UPNA (Ariz et al., 2016) and UPNA-Synth (Larumbe et al., 2017), containing 8,000 and 9,000 images, respectively. Both datasets have been pre-processed and categorised as described in section III A. Table 3 shows that a similar level of accuracy was achieved on the UPNA and UPNA-Synth datasets (90.73 % and 87.01 %, respectively) as compared to the 89.79 % achieved on the reserved evaluation set of the BIWI dataset as provided in Table 2.

In addition to accuracy, Tables 2 and Table 3 also report the sensitivity and specificity evaluation metrics. Sensitivity evaluates the models' ability to predict true positives for each category, while specificity evaluates the models' ability to predict true negatives for each category. As shown in Tables 2 and 3, sensitivity is above 98 % when the models are evaluated on datasets containing all nine categories, suggesting that the models can accurately predict true positives for all categories. Similarly, over 80 % specificity is achieved on all three datasets, suggesting that the models can correctly predict false negatives for all categories.

Fig. 5 shows the specificity, sensitivity, and accuracy of the vertical, horizontal, and combined models. The performance of the models is consistent mainly, with some minor differences. Firstly, the horizontal model slightly outperforms the vertical model, likely due to the more significant differentiation in pose appearance. For example, as a person turns their head to one side, only one eye and ear may be visible, whereas both are visible when facing the centre. This differentiation is not available for the vertical classes as a person looking centre will show all facial features regardless of whether they look up or down. Secondly, the combined model has decreased accuracy and specificity due to the nature of the approach. For the combined model to correctly classify, both the horizontal and vertical models must give correct classifications. Therefore, the combined models' performance is limited to the weakest performance of horizontal or vertical models.

4.2. Head roll estimation (Exp 1_c)

The aforementioned HRE solution was evaluated using the entire BIWI dataset (Fanelli et al., 2013), comprising approximately 15,678 images. Overall, the method achieved an MAE of 9.91°. Table 4 displays the MAE of this approach when different maximum yaw values are applied to filter the images used for evaluation; the yaw range stops at 85°, as this includes all instances in the BIWI dataset. The results suggest that the accuracy of this method is affected by varying the yaw of the head; the higher the yaw, the less accurate the roll estimation.

Table 5 compares a range of HRE methods which provide Euler angles. Each method, including ours, is evaluated on the BIWI dataset (Fanelli et al., 2013) limited to ± 99° as evaluated in (Asperti and Filippini, 2023). Other than our proposed method, the remaining methods in Table 5 use deep learning for HRE. Despite not producing state-of-the-art performance, Table 5 suggests that the performance of our proposed HRE methods is comparable to the deep learning methods. However, unlike the works presented in Table 5, our approach does not use deep learning to calculate head roll. Therefore, it is likely that our method may be computationally less expensive. However, this assumes that the required key points for our head roll calculation may be performed without deep learning.

Table 2
Results of the discrete hpe on the reserved biwi evaluation set (20%).

Category	Accuracy	Sensitivity	Specificity
Vertical (3 Categories)	89 %	93.4 %	90.3 %
Horizontal (3 Categories)	92.8 %	94.5 %	89.9 %
Combined (9 Categories)	86.8 %	98.1 %	80.6 %

Table 3
Cross-dataset validation for the proposed discrete hpe.

Category	Accuracy	Sensitivity	Specificity
UPNA	90.7 %	98.6 %	82.9 %
UPNA-Synth	87 %	98.1 %	75.9 %

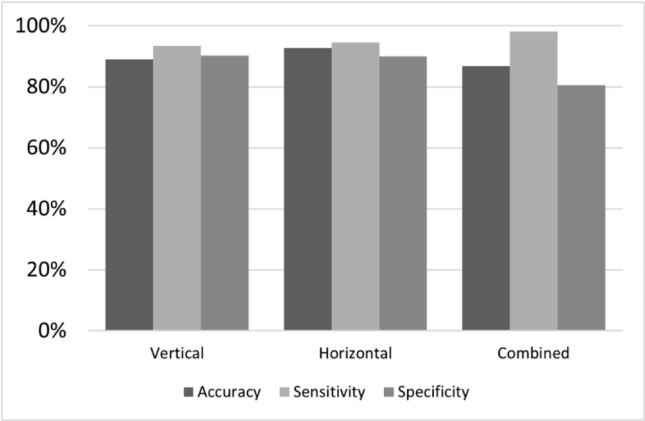


Fig. 5. Performance comparison of the proposed HPE including accuracy, sensitivity, and accuracy for the three HPE experiments: horizontal, vertical, and combined.

Table 4
Head Roll estimation results with a varying range of maximum yaw limits.

Maximum Yaw°	Total Images	MAE°
5	415	2.1
10	1396	2.4
15	2891	2.9
20	4451	3.5
25	6022	4.2
30	7455	5.2
35	8916	6.1
40	10,289	6.9
45	11,471	7.5
50	12,649	8.1
55	13,764	8.7
60	14,510	9.2
65	15,077	9.5
70	15,463	9.7
75	15,613	9.9
80	15,665	9.9
85	15,678	9.9

4.3. Iris detector (Exp 2_a, Exp 2_b)

The iris detector model achieved an mAP of 99.5 % when evaluated using the proposed iris detection dataset described in Section III-B-1. As described in Section III-B-2, experiments included using a YOLO v5 model and a Mask R-CNN. Table 6 suggests that the performance of both models is similar. However, the YOLO v5 model indicates slightly improved performance. Further training and validation performances of proposed iris detection are shown in Figs. S2 and S3 in Supplementary material.

Furthermore, Mask R-CNN averaged a mAP score of 78 % across ranges of 0.5 to 0.9 IoU. On a similar scale of 0.5 to 0.95 IoU, the Yolo v5 model achieved an average mAP of 77 %. Note that the Yolo v5 model's average is inclusive of up to 0.95 IoU, which should go against the Yolo v5 model's score, yet it performs similar to the Mask RCNN model (with 78 % mAP). Fig. 6 shows a sample of iris detections from a custom-trained Yolov5 model for unseen samples from the test set. It can be noticed that the proposed iris detection can perform in challenging cases with varying illuminations, orientations, sizes, and other diversities. This demonstrates the generalisation of the proposed iris detector, which would be useful for EGE, particularly in realistic environments.

Table 5
A Comparison of head roll estimation methods evaluated using the BIWI dataset with the associated MAE.

Paper	MAE°
Ours	9.9
(Kumar et al., 2017)	16.2
(Zhu et al., 2016)	8.8
(Bulat and Tzimiropoulos, 2017)	7.6
(Fanelli et al., 2011)	8.9
(Kazemi and Sullivan, 2014)	23.1
(Basak et al., 2021)	9.8

Table 6
Performance of the proposed iris Detection method validated over unseen instances (IoU:0.5) (Lin, et al., 2014).

Model	mAP (0.5)	Recall	Precision
YOLO v5	99.5 %	99.6 %	99.6 %
Mask R-CNN	99.0 %	99.5 %	98.9 %

4.4. Eye state estimator (Exp 3)

The ESE model achieved 98 % validation accuracy when evaluated over the reserved 20 % of the MRL dataset (with the remaining 80 % used for training). Table 7 presents additional statistics for both the open and closed categories, in addition to the macro average (i.e., computed without considering the class proportions) and weighted average (i.e., computed assuming the class proportions) of the categories. These results further show that each category's precision, recall, and F1-Score are high, between 98 % and 99 %. Further training and validation performance for the ESE is shown in Fig. S1 of Supplementary material.

Overall, the ESE model achieved 92 % accuracy on the unseen mEBAL dataset. This is lower than that of the validation subset of the MRL dataset. However, some difference is expected between a validation subset and a completely unseen dataset. Table 8 presents additional statistics for each of the eye states. As shown in Table 8, the precision for the open state is higher than that of the closed state, 99 % and 90 %, respectively. Similar patterns can be seen in the recall and F1-Score results. This suggests that the model is more likely to misidentify a closed eye as an open eye. However, the results remain relatively high, with an average accuracy of 92 %, as reported previously.

5. Discussions

The results described in section IV suggest that the two-model approach to categorical HPE provides significantly higher accuracy than the single-model approach. However, one limitation of the two-model system is that the combined model (for nine categories of head pose) cannot produce an accuracy higher than that of the lowest performing individual model (i.e., horizontal and vertical models), in this case, the vertical model. However, this can be resolved by training more accurate models, beginning by collecting further data instances for all categories.

Accuracies of 86.79 %, 90.73 %, and 87.01 % are reported for the 9-category HPE task when evaluating the models using the BIWI, UPNA and UPNA-Synth datasets, respectively. To the best of the authors' knowledge, this is the first attempt at predicting nine head pose categories; other works, such as (Elharrouss et al., 2020), predict only three categories (left, centre, and right). Furthermore, the categorisation (i.e., classification) of HPE would enable the extraction of discrete-level facial movements (in the form of facial encoding), which would be useful for comprehensive psychological profiling tasks along with other applications. To achieve this, two models have been combined, the first for predicting horizontal classes (left, centre, and right) and the second for predicting vertical classes (up, centre, and down), to provide the nine categories. Section IV-A suggests that the two-model approach produced significantly higher than the single-model, which achieved only 70.1 % on the BIWI dataset compared to the 89.04 % achieved using the two-model system on the same dataset.

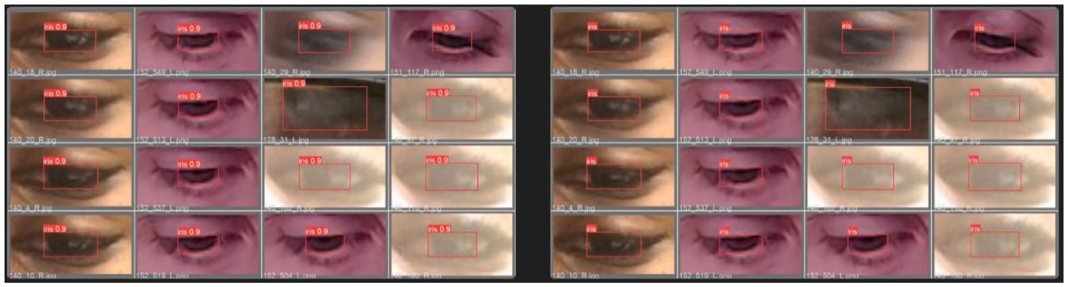


Fig. 6. Yolo prediction sample from unseen test set. To the left, we can see the box predictions of the highest confidence matching the final predictions on the right (for the purely unseen samples).

Table 7
Results of the eye state estimation on the unseen 20% reserved subset of the mrl dataset (Fusek, 2018)test set.

State	Precision	Recall	F1-Score
Open	98.3 %	98.2 %	98.2 %
Closed	98.1 %	99.5 %	98.3 %
Macro Average	98.3 %	98 %	98.1 %
Weighted Average	98.4 %	98.1 %	98.2 %

Table 8
Results of the validation of the eye state estimation on the additional unseen mebal dataset (Daza et al., 2020).

State	Precision	Recall	F1-Score
Open	99 %	98.5 %	94.1 %
Closed	90.3 %	90.1 %	88.2 %
Macro Average	90.2 %	98.5 %	91.5 %
Weighted Average	93.5 %	92.3 %	91.7 %

The proposed HRE method performs a single calculation based on two key points. It is possible to perform the same calculation using a variety of other key points that can be identified in the face, as provided by the facial landmark identification methods (e.g., DLib); this provides two opportunities for future work. Firstly, alternative key points may be used when the desired key points cannot be located. Secondly, it may be beneficial to the accuracy of the estimation to perform several calculations and then take the average of all measurements. Further work will aim to implement and evaluate this proposal.

The eye tracking experiments reported slightly improved (+0.5 mAP) performance when using a YOLO v5 model. However, a similar performance was achieved using Mask R-CNN. However, it is important to note that the YOLOv5 has advantages in terms of processing speeds; it is approximately 2.5x faster than Mask R-CNN at object detection tasks such as iris detection (Fang et al., 2021). This enables the use of the proposed method in realistic situations (e.g., interrogations, cognitive tasks requiring real-time analysis) with the ability to perform the facial analysis in real time.

The outcomes from HPE detectors, ESE, and EGE are integrated into the proposed method in composition with the Dlib landmarks identification to encode the real-time video stream and image/s to produce a comprehensive profile comprising head pose, eye movements, and eye state information. The complete cycle of the proposed DeFaP is described in Algorithm 1. The outcomes have been generated for the given datasets using proposed HPE and eye-tracking DL models. Regarding HPE, the categorical HPE performance results achieved are provided in Tables 2 and 3, and the head roll performance is described in Table 4. Moreover, regarding eye tracking, the iris tracking performance is described in Table 6, and the ESE performance is described in Tables 6 and 7.

Furthermore, the outcomes suggest that the proposed system improves in several ways compared to the existing literature. For example, the data extractor in Fathom (Buckingham et al., 2015) is unreliable. Potentially, this could cause significant issues regarding generalisation when using these extracted features. Moreover, varying lighting conditions would highly affect eye estimation. Specifically, it is unclear how an eye state is classified, labelled, and annotated and how difficult the classification of an eye microstate would be with conventional machine learning models without using DL, landmarks, etc. Additionally, the literature on the error of individual classifiers is unclear, for example, regarding the features used (i.e., accuracy and reliability) or the validation method. Likewise, the Silent Talker system (Rothwell et al., 2006) is affected by similar limitations. Additionally, some limitations surround noisy outputs generated from the system, which may result in inappropriate analytical outcomes.

Moreover, many works, such as (Fuhl et al., 2016; Santini et al., 2018), rely on pupil estimation for EGE. However, the literature suggests that iris detection may be more reliable than pupil detection due to the increased colour contrast and larger size (compared to pupil) (Sigut and Sidha, 2011). Therefore, the proposed DeFaP method may provide more reliable EGE due to its use of iris detection, particularly under varying brightness levels. Furthermore, each aspect of DeFaP is validated over multiple standard datasets, unlike works such as (Rothwell et al., 2006; Buckingham et al., 2015; Fuhl et al., 2016; Santini et al., 2018; Liu et al., 2021), which either do not cross-validate (on different datasets) or have not reported the results of this.

Table 9 compares the related head and eye pose and state estimation methods. It suggests that only our proposed DeFaP method provides HPE, EGE, ESE, classification (in the form of detector array), and regression (for HRE) and is evaluated using cross-dataset validation (i.e., using multiple unseen datasets to validate the performance). Furthermore, unlike the proposed DeFaP, several works such as (Patacchiola and Cangelosi, 2017; Santini et al., 2018; Ruiz et al., 2018; Santini et al., 2018; Liu et al., 2021) are not evaluated using cross-dataset validation. Therefore, it is difficult to assess their generalisation. Many solutions, such as (Elharrouss et al., 2020; Sreekanth et al., 2018), provide only HPE and do not provide EGE or ESE. Conversely, some EGE estimation methods, such as (Yiu, 2019; Santini et al., 2018; Sanyal and Chakrabarty, 2019; Fuhl et al., 2016; Santini et al., 2018; Liu et al., 2021), provide only EGE or ESE and do not provide HPE. Moreover, many HPE works, such as (Patacchiola and Cangelosi, 2017; Ruiz et al., 2018) provide only continuous data, which is not easily human-understandable, unlike discrete labelled classes in the proposed DeFaP. In (Elharrouss et al., 2020) and (Sreekanth et al., 2018), a similar approach to DeFaP's discrete HPE classification is described; however, only horizontal categories are classified, producing three classes compared to the 9 in DeFaP. Moreover, as mentioned previously, neither EGE nor ESE is provided, unlike the proposed DeFaP method. In addition, we propose eye-tracking based on iris detection, which is a larger eyeball segment with no dilation and might be able to tolerate slight noise in the background. More importantly, the varying dilations within the pupil will not affect the iris localisation and, therefore, eye tracking performance.

Finally, we report some limitations of the proposed DeFaP approach,

Table 9
A comparison of related head and eye pose and state estimation systems, including the provision of HPE, EGE, ESE, classification data (if classification is performed, the number of categories for classification tasks), continuous data, and cross-dataset validation.

Paper	HPE	EGE	ESE	Classification	No. Classes	Regression	Cross-Dataset Validation
Ours	✓	✓	✓	✓	**9**	✓	✓
(Elharrouss et al., 2020)	✓	×	×	✓	3	×	✓
(Sreekanth et al., 2018)	✓	×	×	✓	3	×	✓
(Ruiz et al., 2018)	✓	×	×	×	0	✓	×
(Patacchiola and Cangelosi, 2017)	✓	×	×	×	0	✓	×
(Sanyal and Chakrabarty, 2019)	×	✓	✓	×	0	✓	✓
(Yiu, 2019)	×	✓	✓	×	0	✓	✓
(Fuhl et al., 2016)	×	✓	×	×	0	✓	✓
(Santini et al., 2018)	×	✓	×	×	0	✓	×
(Santini et al., 2018)	×	✓	×	×	0	✓	×
(Liu et al., 2021)	×	✓	×	×	0	✓	×

which can be addressed in future works. Firstly, the proposed custom-trained models could be encoded as a complete application with public access. This would help a wide range of research and development communities with diverse applications. Secondly, similar to HPE, an alternative approach for the proposed iris detection model could be implemented to automatically categorise the eye gaze into nine states (e.g., looking left, right, top-left, etc.). This could then be compared with the proposed iris-detection-based eye gaze estimation.

6. Conclusion and future work

Inspired by our previous research (Khan et al., 2021), this study presented an explicit, non-restrained, non-invasive facial and gaze comprehension approach, including categorical HPE (9 categories), a simple head roll calculation method, *EGE*, and ESE. As reported in Section V, the methods have been evaluated using several standard head and eye pose estimation datasets, indicating state-of-the-art performance. Moreover, using a categorical classification approach to HPE produces a more human-understandable output than a geometric output. The approach could be extended to the animal world to evaluate animal psychological and cognitive behaviours further.

Furthermore, the precision and reliability of head pose and gaze estimation are vital to comprehending various situations, activities, and psychological or cognitive tasks, for example, driver behaviour analysis, student attention analysis, investigation-based tasks, and other highly impactful applications in healthcare (e.g., autism, animal welfare etc.). The proposed DeFaP approach is likely more reliable than previous works due to its core architecture (utilising DTL) and cross-dataset validation, as reported in Section IV. Moreover, the increased human understandability provided by DeFaP (i.e., discrete head pose categories) will broaden the range of domains and applications the technology can impact, as an understanding of 3D geometry is not required of the user. For instance, using the example of driver analysis, DeFaP would provide human-understandable head positions, which may be used to identify if the driver is regularly checking the mirrors; EGE will allow users to determine if the driver is looking ahead at the road or elsewhere, and ESE will allow the evaluation of the drivers level of alertness (i.e., awake, drowsy, or asleep).

Multiple aspects of future work are planned for this project. For example, the proposed DeFaP algorithm (Algorithm 1) can be implemented to provide a comprehensive head pose and eye-tracking application. Currently, all trained models are available, so implementation is relatively trivial. Furthermore, future work will demonstrate the implementation of the DeFaP tool in several domains and applications, such as driver attention analysis, deception detection system, and other high-impact healthcare applications (such as in autism).

CRediT authorship contribution statement

Wasiq Khan: Conceptualization, Methodology, Software, Data curation, Visualization, Formal analysis, Writing – original draft, Supervision. **Luke Topham:** Writing – original draft, Data curation, Methodology, Formal analysis. **Hiba Alsmadi:** Software, Writing – review & editing. **Ala Al Kafri:** Validation, Writing – review & editing, Visualization. **Hoshang Kolivand:** Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2024.124425.

References

Mukherjee, S. S., & Robertson, N. M. (2015). Deep Head Pose: Gaze-Direction Estimation in Multimodal Video. *IEEE Transactions on Multimedia, 17*(11), 2094–2107. https://doi.org/10.1109/TMM.2015.2482819

W. Khan, K. Crockett, J. O'Shea, A. Hussain, and B. M. Khan, "Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection," *Expert Systems with Applications*, vol. 169, no. February 2020, p. 114341, 2021, doi: 10.1016/j.eswa.2020.114341.

Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261–8265).

Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. (2018). Near Real-Time Comprehension Classification with Artificial Neural Networks: Decoding e-Learner Non-Verbal Behavior. *IEEE Transactions on Learning Technologies, 11*(1), 5–12. https://doi.org/10.1109/TLT.2017.2754497

Venturelli, M., Borghi, G., Vezzano, R., & Cucchiara, R. (2016). Deep Head Pose Estimation from Depth Data for In-Car Automotive Applications. In *International Workshop on Understanding Human Activities through 3D Sensors* (pp. 74–85).

Mittal, A., Kumar, K., Dhamija, S., & Kaur, M. (2016). Head movement-based driver drowsiness detection: A review of state-of-art techniques. In *Proceedings of 2nd IEEE International Conference on Engineering and Technology* (pp. 903–908). https://doi.org/10.1109/ICETECH.2016.7569378

Kong, S. G., & Mbouna, R. O. (2015). Head Pose Estimation from a 2D Face Image Using 3D Face Morphing With Depth Parameters. *IEEE Transactions on Image Processing, 24*(6), 1801–1808. https://doi.org/10.1109/TIP.2015.2405483

Huang, Z., Zhang, J., & Shan, H. (2023). When Age-Invariant Face Recognition Meets Face Age Synthesis: A Multi-Task Learning Framework and a New Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(6), 7917–7932. https://doi.org/10.1109/TPAMI.2022.3217882

Baxter, R. H., Leach, M. J. V., Mukherjee, S. S., & Robertson, N. M. (2015). An adaptive motion model for person tracking with instantaneous head-pose features. *IEEE Signal Processing Letters, 22*(5), 578–582. https://doi.org/10.1109/LSP.2014.2364458

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(4), 607–626. https://doi.org/10.1109/TPAMI.2008.106

Kumar, A., Kaur, A., & Kumar, M. (2019). Face detection techniques: A review. *Artificial Intelligence Review, 52*(2), 927–948. https://doi.org/10.1007/s10462-018-9650-2

Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K., & Wang, J. (2022). MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation. *IEEE Transactions on Multimedia, 24*, 2449–2460. https://doi.org/10.1109/TMM.2021.3081873

Liu, H., Nie, H., Zhang, Z., & Li, Y. F. (2021). Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction. *Neurocomputing, 433*, 310–322. https://doi.org/10.1016/j.neucom.2020.09.068

Patacchiola, M., & Cangelosi, A. (2017). Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods. *Pattern Recognition, 71*, 132–143. https://doi.org/10.1016/j.patcog.2017.06.009

Liu, H., Liu, T., Chen, Y., Zhang, Z., & Li, Y. F. (2022). EHPE: Skeleton Cues-based Gaussian Coordinate Encoding for Efficient Human Pose Estimation. *IEEE Transactions on Multimedia, PP*, 1–12. https://doi.org/10.1109/TMM.2022.3197364

Liu, H., Zhang, C., Deng, Y., Liu, T., Zhang, Z., & Li, Y. F. (2023). Orientation Cues-Aware Facial Relationship Representation for Head Pose Estimation via Transformer. *IEEE Transactions on Image Processing, 32*, 6289–6302. https://doi.org/10.1109/TIP.2023.3331309

Diaz-Chito, K., Hernández-Sabaté, A., & López, A. M. (2016). A reduced feature set for driver head pose estimation. *Applied Soft Computing Journal, 45*, 98–107. https://doi.org/10.1016/j.asoc.2016.04.027

G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz, "Robust model-based 3D head pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, pp. 3649–3657, doi: 10.1109/ICCV.2015.416.

H. Moon and M. L. Miller, "Estimating Facial Pose from a Sparse Representation," 2009.

Rothwell, J., Bandar, Z., O'Shea, J., & McLean, D. (2006). Silent talker: A new computer-based system for the analysis of facial cues to deception. *Applied Cognitive Psychology, 20*(6), 757–777. https://doi.org/10.1002/acp.1204

Ma, B., Chai, X., & Wang, T. (2013). A novel feature descriptor based on biologically inspired feature for head pose estimation. *Neurocomputing, 115*, 1–10. https://doi.org/10.1016/j.neucom.2012.11.005

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research, 10*, 1755–1758.

G. Bradski, "OpenCV Library," 2000. opencv.org (accessed Oct. 16, 2023).

Q. Liu, J. Yang, J. Deng, and K. Zhang, "Robust facial landmark tracking via cascade regression," *Pattern Recognition*, vol. 66, no. December 2016, pp. 53–62, 2017, doi: 10.1016/j.patcog.2016.12.024.

Nanduri, A., & Park, C. (2024). Semi-supervised Cross-Spectral Face Recognition with Small Datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 588–596).

Himmi, S., Parret, V., Chhatkuli, A., & Van Gool, L. (2024). MS-EVS : Multispectral event-based vision for deep learning based face detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 616–625).

P. Yin and W. Yu, "Nonlinear dynamical system iteration applied in video face feature extraction and recognition," *Evolutionary Systematics*, no. 0123456789, 2024, doi: 10.1007/s12530-023-09562-5.

Zeebaree, I. M., & Kareem, O. S. (2023). Face Mask Detection Using Haar Cascades Classifier To Reduce The Risk Of Coved-19. *International Journal of Mathematics and Computer Science, 2*, 19–27. https://doi.org/10.59543/ijmscs.v2i.7845

N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018, vol. 2018-June, pp. 2074–2083, doi: 10.1109/CVPRW.2018.00281.

T. Y. Yang, Y. T. Chen, Y. Y. Lin, and Y. Y. Chuang, "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, vol. 2019-June, pp. 1087–1096, doi: 10.1109/CVPR.2019.00118.

Hong, C., Yu, J., Zhang, J., Jin, X., & Lee, K. H. (2019). Multimodal Face-Pose Estimation With Multitask Manifold Deep Learning. *IEEE Transactions on Industrial Informatics, 15*(7), 3952–3961. https://doi.org/10.1109/TII.2018.2884211

Ranjan, R., Patel, V. M., & Chellappa, R. (2019). HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(1), 121–135. https://doi.org/10.1109/TPAMI.2017.2781233

Jyotsna, C., Amudha, J., Ram, A., & Nollo, G. (2022). IntelEye: An Intelligent Tool for the Detection of Stressful State based on Eye Gaze Data while Watching Video. *Procedia Computer Science, 218*, 1270–1279. https://doi.org/10.1016/j.procs.2023.01.105

Clark, J. R., Stanton, N. A., & Revell, K. M. A. (2019). Directability, eye-gaze, and the usage of visual displays during an automated vehicle handover task. *Transportation Research, 67*, 29–42. https://doi.org/10.1016/j.trf.2019.10.005

Khan, M. Q., & Lee, S. (2019). Gaze and eye tracking: Techniques and applications in ADAS. *Sensors, 19*(24), pp. https://doi.org/10.3390/s19245540

L. R. D. Murthy and P. Biswas, "Deep Learning-based Eye Gaze Estimation for Military Aviation," in *IEEE Aerospace Conference Proceedings*, 2022, vol. 2022-March, pp. 1–8, doi: 10.1109/AERO53065.2022.9843506.

Amer, S. G., Kamh, S. A., Elshahed, M. A., & Ramadan, R. A. (2021). Wheelchair Control System based Eye Gaze. *International Journal of Advanced Computer Science and Applications, 12*(6), 895–900. https://doi.org/10.14569/IJACSA.2021.01206104

Drouard, V., Horaud, R., Deleforge, A., Ba, S., & Evangelidis, G. (2017). Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regressions. *IEEE Transactions on Image Processing, 26*(3), 1428–1440. https://doi.org/10.1109/TIP.2017.2654165

Khosravan, N., Celik, H., Turkbey, B., Jones, E. C., Wood, B., & Bagci, U. (2019). A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. *Medical Image Analysis, 51*, 101–115. https://doi.org/10.1016/j.media.2018.10.010

Yiu, Y. H., et al. (2019). DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of Neuroscience Methods, 324*(May), 108307. https://doi.org/10.1016/j.jneumeth.2019.05.016

Medeiros, P. A. de L., et al. (2022). Efficient machine learning approach for volunteer eye-blink detection in real-time using webcam. *Expert Systems with Applications, 188*, Article 116073. https://doi.org/10.1016/j.eswa.2021.116073

Zemblys, R., Niehorster, D. C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods, 50*(1), 160–181. https://doi.org/10.3758/s13428-017-0860-3

S. Hickson, V. Kwatra, N. Dufour, A. Sud, and I. Essa, "Eyemotion: Classifying facial expressions in VR using eye-tracking cameras," in *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, 2019, pp. 1626–1635, doi: 10.1109/WACV.2019.00178.

L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11218 LNCS, no. Dl, pp. 625–642, doi: 10.1007/978-3-030-01264-9_37.

P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, vol. 2019-Octob, pp. 6911–6920, doi: 10.1109/ICCV.2019.00701.

Jordan, A. A., Pegatoquet, A., Castagnetti, A., Raybaut, J., & Le Coz, P. (2020). Deep Learning for Eye Blink Detection Implemented at the Edge. *IEEE Embedded Systems Letters, 13*(3), 130–133. https://doi.org/10.1109/LES.2020.3029313

Y. Yu and J. M. Odobez, "Unsupervised representation learning for gaze estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, no. ii, pp. 7312–7322, doi: 10.1109/CVPR42600.2020.00734.

Santini, T., Fuhl, W., & Kasneci, E. (2018). PuReST: Robust pupil tracking for real-time pervasive eye tracking. *Eye Tracking Research and Applications Symposium (ETRA)*. https://doi.org/10.1145/3204493.3204578

Khan, W., Hussain, A., Kuru, K., & Al-Askar, H. (2020). Pupil localisation and eye centre estimation using machine learning and computer vision. *Sensors, 20*(13), 1–18. https://doi.org/10.3390/s20133785

J. H. Choi, K. Il Lee, Y. C. Kim, and B. C. Song, "Accurate eye pupil localization using heterogenous CNN models," *2019 IEEE Int. Conf. Image Process.*, pp. 2179–2183, 2019.

Rakhmatulin, I., & Duchowski, A. T. (2020). Deep neural networks for low-cost eye tracking. *Procedia Computer Science, 176*, 685–694. https://doi.org/10.1016/j.procs.2020.09.041

K. Krafka *et al.*, "Eye Tracking for Everyone," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 2176–2184, doi: 10.1109/CVPR.2016.239.

S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11217 LNCS, pp. 741–757, doi: 10.1007/978-3-030-01261-8_44.

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2019). MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(1), 162–175. https://doi.org/10.1109/TPAMI.2017.2778103

T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 334–352.

Jayanthi, J., Lydia, E. L., Krishnaraj, N., Jayasankar, T., Babu, R. L., & Suji, R. A. (2021). An effective deep learning features based integrated framework for iris detection and recognition. *Journal of Ambient Intelligence and Humanized Computing, 12*(3), 3271–3281. https://doi.org/10.1007/s12652-020-02172-y

E. Severo *et al.*, "A Benchmark for Iris Location and a Deep Learning Detector Evaluation," in *Proceedings of the International Joint Conference on Neural Networks*, 2018, vol. 2018-July, pp. 1–7, doi: 10.1109/IJCNN.2018.8489638.

J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018. [Online]. Available: http://arxiv.org/abs/1804.02767.

Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *NeuroImage, 20*(1), 114–124. https://doi.org/10.1016/S1053-8119(03)00298-2

B. R. Ibrahim *et al.*, "Embedded System for Eye Blink Detection Using Machine Learning Technique," in *1st Babylon International Conference on Information Technology and Science 2021, BICITS 2021*, 2021, vol. 2021, no. Bicits, pp. 58–62, doi: 10.1109/BICITS51482.2021.9509908.

Hu, G., et al. (2020). Towards Real-Time Eyeblink Detection in the Wild: Dataset, Theory and Practices. *IEEE Transactions on Information Forensics and Security, 15*, 2194–2208. https://doi.org/10.1109/TIFS.2019.2959978

R. Sanyal and K. Chakrabarty, "Two Stream Deep Convolutional Neural Network for Eye State Recognition and Blink Detection," in *2019 3rd International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2019*, 2019, doi: 10.1109/IEMENTech48150.2019.8981102.

Fogelton, A., & Benesova, W. (2018). Eye blink completeness detection. *Computer Vision and Image Understanding, 176–177*(September), 78–85. https://doi.org/10.1016/j.cviu.2018.09.006

Ryan, C., et al. (2021). Real-time face & eye tracking and blink detection using event cameras. *Neural Networks, 141*(2021), 87–97. https://doi.org/10.1016/j.neunet.2021.03.019

Fanelli, G., Dantone, M., Gall, J., Fossati, A., & Van Gool, L. (2013). Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision, 101*(3), 437–458. https://doi.org/10.1007/s11263-012-0549-0

Ariz, M., Bengoechea, J. J., Villanueva, A., & Cabeza, R. (2016). A novel 2D/3D database with automatic face annotation for head tracking and pose estimation. *Computer Vision and Image Understanding, 148*, 201–210. https://doi.org/10.1016/j.cviu.2015.04.009

Larumbe, A., Ariz, M., Bengoechea, J. J., Segura, R., Cabeza, R., & Villanueva, A. (2017). Improved Strategies for HPE Employing Learning-by-Synthesis Approaches. In *IEEE International Conference on Computer Vision Workshops* (pp. 1545–1554). https://doi.org/10.1109/ICCVW.2017.182

Google, "MediaPipe Face Mesh," *GitHub*, 2020. https://google.github.io/mediapipe/solutions/face_mesh.html (accessed Feb. 22, 2022).

Neill, H. (2018). Trigonometry: A complete introduction. *Möbius*.

B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," *UIST 2013 - Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, pp. 271–280, 2013, doi: 10.1145/2501988.2501994.

Villanueva, A., Ponz, V., Sesma-Sanchez, L., Ariz, M., Porta, S., & Cabeza, R. (2013). Hybrid method based on topography for robust detection of iris center and eye corners. *ACM Transactions on Multimedia Computing, Communications, and Applications, 9*(4), 1–20. https://doi.org/10.1145/2501643.2501647

W. Khan, A. Hussain, B. M. Khan, and K. Crockett, "Outdoor mobility aid for people with visual impairment: Obstacle detection and responsive framework for the scene perception during the outdoor mobility of people with visual impairment," *Expert Systems with Applications*, vol. 228, no. August 2022, p. 120464, 2023, doi: 10.1016/j.eswa.2023.120464.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

Yang, W., Li, S., Ouyang, W., Li, H., & Wang, X. (2017). Learning feature pyramids for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1281–1290). https://doi.org/10.4028/www.scientific.net/AMR.988.290

Plastiras, G., Kyrkou, C., & Theocharides, T. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *ACM International Conference Proceeding Series* (pp. 779–788). https://doi.org/10.1145/3243394.3243692

Cao, Z., Liao, T., Song, W., Chen, Z., & Li, C. (2021). Detecting the shuttlecock for a badminton robot: A YOLO based approach. *Expert Systems with Applications, 164*. https://doi.org/10.1016/j.eswa.2020.113833

Fusek, R. (2018). Pupil Localization Using Geodesic Distance. *Lecture Notes in Computer Science Book Series*.

Daza, R., Morales, A., Fierrez, J., & Tolosana, R. (2020). MEBAL: A multimodal database for eye blink detection and attention level estimation. *International Conference on Multimodal Interaction*, 32–36. https://doi.org/10.1145/3395035.3425257

Jung, Y., Kim, D., Son, B., & Kim, J. (2017). An eye detection method robust to eyeglasses for mobile iris recognition. *Expert Systems with Applications, 67*, 178–188. https://doi.org/10.1016/j.eswa.2016.09.036

O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "LFR face dataset:Left-Front-Right dataset for pose-invariant face recognition in the wild," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies, ICIoT 2020*, 2020, pp. 124–130, doi: 10.1109/ICIoT48696.2020.9089530.

A. Asperti and D. Filippini, *Deep Learning for Head Pose Estimation: A Survey*, vol. 4, no. 4. Springer Nature Singapore, 2023.

A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors," in *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, 2017, pp. 258–265, doi: 10.1109/FG.2017.149.

Zhu, X., Liu, X., Lei, Z., Li, S. Z., & Shi, H. (2016). Face Alignment Across Large Poses A 3D Solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 146–155).

A. Bulat and G. Tzimiropoulos, "How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 1021–1030, doi: 10.1109/ICCV.2017.116.

Fanelli, G., Thibaut, W., Gall, J., & Van Gool, L. (2011). Real time head pose estimation from consumer depth cameras. *Joint pattern recognition symposium*.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1867–1874). https://doi.org/10.1109/CVPR.2014.241

Basak, S., Corcoran, P., Khan, F., McDonnell, R., & Schukat, M. (2021). Learning 3D head pose from synthetic data: A semi-supervised approach. *IEEE Access, 9*, 37557–37573. https://doi.org/10.1109/ACCESS.2021.3063884

T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, vol. 8693 LNCS, no. PART 5, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

Fang, Y., Guo, X., Chen, K., Zhou, Z., & Ye, Q. (2021). Accurate and Automated Detection of Surface Knots on Sawn Timbers Using YOLO-V5 Model. *BioResources, 16*(3), 5390–5406. https://doi.org/10.15376/biores.16.3.5390-5406

Buckingham, F. J., Crockett, K. A., Bandar, Z. A., & O'Shea, J. D. (2015). FATHOM: A neural network-based non-verbal human comprehension detection system for learning environments. In *IEEE SSCI 2014–2014 IEEE Symposium Series on Computational Intelligence – CIDM 2014: 2014 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 403–409). https://doi.org/10.1109/CIDM.2014.7008696

P. Sreekanth, U. Kulkarni, S. Shetty, and S. M. Meena, "Head Pose Estimation using Transfer Learning," in *Proceedings of the 2018 International Conference on Recent Trends in Advanced Computing, ICRTAC-CPS 2018*, 2019, pp. 73–79, doi: 10.1109/ICRTAC.2018.8679209.

N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2018-June, pp. 2155–2164, 2018, doi: 10.1109/CVPRW.2018.00281.

Fuhl, W., Tonsen, M., Bulling, A., & Kasneci, E. (2016). Pupil detection for head-mounted eye tracking in the wild: An evaluation of the state of the art. *Machine Vision and Applications, 27*(8), 1275–1288. https://doi.org/10.1007/s00138-016-0776-4

T. Santini, W. Fuhl, and E. Kasneci, "PuRe: Robust pupil detection for real-time pervasive eye tracking," *Comput. Vis. Image Underst.*, vol. 170, no. December 2017, pp. 40–50, 2018, doi: 10.1016/j.cviu.2018.02.002.

Liu, M., Li, Y., & Liu, H. (2021). Robust 3-D Gaze Estimation via Data Optimization and Saliency Aggregation for Mobile Eye-Tracking Systems. *IEEE Transactions on Instrumentation and Measurement, 70*. https://doi.org/10.1109/TIM.2021.3065437

Sigut, J., & Sidha, S. A. (2011). Iris center corneal reflection method for gaze tracking using visible light. *IEEE Transactions on Bio-Medical Engineering, 58*(2), 411–419. https://doi.org/10.1109/TBME.2010.2087330