IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 17, 2021, accepted October 5, 2021, date of publication October 14, 2021, date of current version October 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120098

Attention Span Prediction Using Head-Pose Estimation With Deep Neural Networks

TRIPTI SINGH[1], MOHAN MOHADIKAR[1], SHILPA GITE[2], SHRUTI PATIL[2], BISWAJEET PRADHAN[3,4], AND ABDULLAH ALAMRI[5]

[1]Symbiosis Centre of Applied Artificial Intelligence, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412 115, India
[2]Computer Science Department, Symbiosis Centre of Applied Artificial Intelligence, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412 115, India
[3]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW 2007, Australia
[4]Earth Observation Centre, Institute of Climate Change, University Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia
[5]Department of Geology and Geophysics, College of Science, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding authors: Shilpa Gite (shilpa.gite@sitpune.edu.in) and Biswajeet Pradhan (biswajeet.pradhan@uts.edu.au)

This work was supported in part by the Symbiosis Centre of Applied Artificial Intelligence (SCAAI); in part by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, Australia; and in part by the Researchers Supporting Project through King Saud University, Riyadh, Saudi Arabia, under Grant RSP-2021/14.

ABSTRACT Automated human pose estimation is evolving as an exciting research area in human activity detection. It includes sophisticated applications such as malpractice detection in the examination, distracted driving, gesture detection, etc., and requires robust and reliable pose estimation techniques. These applications help to map the attention of the user with head pose estimation (HPE) metrics supported by emotion and gaze analysis. This paper solves the problem of attention score estimation with HPE. The proposed method ensures ease of implementation while addressing head pose estimation using 68 facial features. Further, to attain reliability and precision, head pose estimation has been implemented as a regression task. The coordinate pair angle method (CPAM) with deep neural network (DNN) regression and elastic net regression is carried out. The use of DNN ensures precision on low lighting, distorted or occluded images. CPAM methodology leverages facial landmark detection and angular difference to estimate head pose. Experimentation results showed that the proposed model could handle large datasets, real-time data processing, significant pose variations, partial occlusions, and diverse facial expressions with a mean absolute error (MAE) of 3° and less. The proposed system was evaluated on three standard databases: the 300W across large poses (300W-LP) dataset, annotated facial landmarks in the wild (AFLW2000) dataset, and the national institute of mental health child emotional faces picture set (NIMH-ChEFS) dataset. The results achieved are on par with recent state-of-the-art methodologies such as anisotropic angle distribution learning (AADL), joint head pose estimation and face alignment algorithm (JFA), rotation axis focused attention network (RAFA-Net), and propose an MAE ranging up to 6°. The paper could achieve remarkable results for attention span prediction using head pose estimation and for many possible future applications.

INDEX TERMS Head pose estimation, CPAM, convolutional neural network, regression, attention span prediction.

I. INTRODUCTION

The need to map the attention span of users has become a necessity in recent decades in almost all spheres of the industry being education [1], [2], medical [3], [4], advertising [5], marketing [6], and many more. To face these real challenges of the digital world, research should be focused on rapid estimates of head pose angles and overcome the problems of lighting conditions [7], blurring [8], occultation [9], or environment conditions [10].

Detecting the malpractices in an e-learning based environment [11], estimating the attention of students in a traditional classroom environment [12], and driver distraction estimation are the recent popular case studies of head pose applications [13].

Recently proposed HPE methods are not robust enough to handle real-world datasets and applications. The proposed dataset testing in a real-world environment infers the

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

diversified collection of images in different environmental conditions and the difference in age, ethnicity, culture, and more. It is claimed that a convolutional neural network (CNN) could be considered one of the best algorithms to work on real-world datasets [14].

Usually, the dataset is assumed to be precise, and the images used for evaluation are well aligned. But practical applications prove these assumptions to be invalid. To overcome these limitations, a coping methodology is proposed in [15], which shows the use of the deep convolutional neural network (DCNN) to handle under-sampling and uncertainty. The proposed method uses dense sampling intervals with multivariate labeling distributions (MLDs) to show input face image head pose angles.

Unstructured behavior of the dynamic environment makes focus-of-attention(FOA) challenging; it gives rise to the possibility of poor quality, occlusion, low-resolution image, disturbance, and non-linear relationship between head pose angle and ground truth value [16]. This research gap encouraged us to incorporate HPE based attention mapping. To make the model generic to multiple applications of head pose estimation [17], proposed applying transfer learning to existing models with minor tweaks. Making the ability of model training a lot easier and more diverse.

Some methods have used embedded attention model systems [18] to enhance the performance of feature expression in multi-level classification with soft stage-wise regression, reducing the number of neurons in each subsequent layer. The study [21] suggests that head pose estimation requires a robust solution to deal with real-world data. The selection of the accurate training dataset mainly contributes to the increased efficiency of the model. Head pose estimation and facial landmark detection have revolutionized the 3D modeling of image datasets [19]–[21].

Head movement has become an essential part of studies involving behavior monitoring. This paper proposed a novel way to estimate head pose for video surveillance and innovative human pose applications using state-of-the-art techniques and map them to give us the user's attention score, which is visualized as a concentration map towards the end of the paper. The proposed methodology was evaluated against classical open-access datasets, namely the 300W-LP dataset, AFLW2000, and NIMH-ChEFS. Finally, the resultant attention map was visualized on the LIRIS Children Spontaneous Facial Expression Video Database.

The main challenges commonly faced in HPE datasets are extreme pose variations, overlapping key points, blending of the head with the environment, video quality deterioration and noise, blurred or occluded images with poor lighting conditions. The state-of-the-art models proposed in the literature take in huge parameters to train data, hence not providing robust solutions.

This research focuses on giving a mathematically inclined lightweight solution to estimate head pose in a real-time environment and produce immediate attention span results over a while. The proposed methodology is robust, can detect faces at a distance, and is compatible with current devices.

This research paper proposes the following significant contributions:

1. Proposes a module to map the user's attention using a robust regression-based model, then merged with CPAM for HPE.
2. The proposition of a model relies on 68 facial landmark locations and estimates head pose through angular difference.
3. A comparative study of the proposed method with traditional regression-based approaches with extensive training parameters provides a sophisticated solution to estimate head pose angles on benchmark datasets. i.e., 300W-LP, AFLW2000, and NIMH-ChEFS.
4. A baseline approach extracts facial keypoint locations and employs ElasticNet and sequential models for regression of pose angles.
5. A model to map the user's concentration level over a time frame and produce an attention score.

The rest of the paper is divided into five sections; namely, section 2 presents related work, section 3 describes the proposed methodology, followed by datasets discussion in Section 4. Section 5 describes experiments and results, and section 6 explains the attention span calculation with a discussion on the findings. Section 7 presents the conclusion and some future research directions.

II. RELATED WORK

Head pose estimation has been a highly researched topic for the past 50 years. It is fascinating how the research has evolved into shaping this study with innumerable methodologies and presented us with several critical findings in this domain.

Head pose estimation can be represented by Euler angles θx, θy, and θz denoting pitch, yaw, and roll, respectively, which is the head's rotation in the X, Y, and Z-axis [21] as presented in figure 1. HPE involves the movement of the jaw and facial muscles in all three degrees of freedom [22].

The review work is divided into five parts based on the research papers studied for this work.

A. REGRESSION-BASED METHODS

Regression is used to map out the relationship between dependent and one or more independent variables. Regression focuses on fitting the regression model on the labeled dataset and predicting pitch, yaw, and roll angles after training, whereas classification focuses on classifying data in discrete bins.

Several studies use regression models for head pose estimation. Many methods have incorporated head pose estimation without any training of neural networks. In [24], the web-based model is combined with the regression model to estimate head pose.

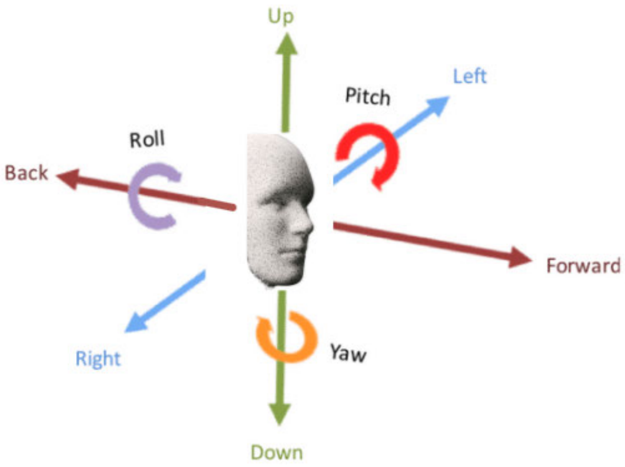


FIGURE 1. Head rotation angles [23].

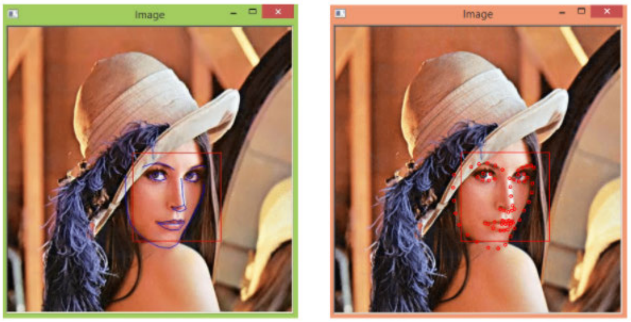


FIGURE 2. Face landmark detection using 68 points model with Dlib [30].

In [25], the model stores facial features in a quad-tree-based representation to estimate HPE angles. The tree data structure is adapted to store a landmark-based representation of face orientation and subdivides it into quadrants to estimate head pose.

B. FACIAL FEATURE RECOMMENDATIONS

Many approaches estimate head pose from facial feature extraction. Head pose estimation and landmark localization are highly correlated. A lot of studies explore the optimization of one concerning another.

In [26], the face is partitioned into seven different facial features, and accurate head pose estimation is done using DCNN, as shown in figure 2. Facial landmarks have also been evaluated through a heatmap generator in a feedforward neural network [27]. The five facial key points of the face, mainly the eyes, ears, and nose, as depicted in figure 3, are processed in 2D soft localization heatmap images. The results are passed to a convolutional neural network to predict the head-pose using regression in [28].

ConvNet model is trained on low-resolution grayscale input images to predict tilt and pan angles without using any facial landmarks in [29], and the proposed methodology is RealHePoNet.

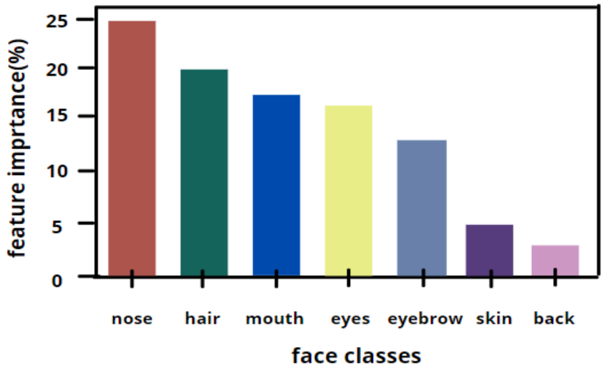


FIGURE 3. Feature importance of face classes [26].

Pose-invariant face recognition (PIFR) proposed in [31] uses a geometric approach for HPE. Frontal face images are constructed using these estimations. In [32], local CNN features for shape and residual pose are predicted by a local neural network (LNet), whereas global neural network (GNet) does the landmark localization for pose estimation in the proposed Joint Head Pose Estimation and Face Alignment Framework (JFA) algorithm. It produces output for both; the head poses estimation and face alignment, exploit the global and local CNN features. More work in this domain suggests combining the HPE angles with Kalman filters, giving high accuracy to handle extreme head poses [33].

In separate work, [34] used a similar approach by proposing an application of cascaded random forest on nine sub-space classifications of the head pose with each specific space trained by a global shape constraint. This classification-based method efficiently handled the problem of significant pose variations.

C. GEOMETRICAL HPE

Some methods incorporate a geometrical and mathematical way of estimating head pose. Anisotropic angle distribution learning (AADL) in [35] suggested that, while increasing fixed central pose and angle interval, the image variations of yaw and pitch angle increased and then started decreasing for yaw angle, creating variations in the two angles. This method worked well on movement caused by blurry images, missing pose images, and occulated images.

Rotation Axis Focused Attention Network (RAFA-Net) [36] explores the significance of spatial structures(fine-grained to coarse) using self-attention layer and concentrated spatial pooling and combines results to form detailed semantic information of a given rotation axis to overcome the limitations of facial landmark localization modules.

Vectors in a rotation matrix are represented as HPE angles and are used to develop a new neural network-based representation. This vector-based approach [37] also suggests using Mean Absolute Error of Vectors (MAEV) to evaluate the precision, as it reflects the actual behavior of profile views better than *MAE*.

FIGURE 4. Poor quality images with occlusion.

D. 3D TRACKING FOR HPE

3D tracking-based methods have gained recent popularity in head pose estimation, mostly in combination with the evaluation of synthetic data and images. Single images can't estimate head pose, as the images need to be mapped using 2D and 3D spaces. In [38], 3D feature points are mapped with depth information to get features from color images using the optimization extended LM (Levenberg-Marquardt) method. This method overcomes the limitations of using RGBD images for direct HPE estimation.

CNN is used to reconstruct the input image of the head to a personalized 3D model, a critical point loss along with Euclidian asymmetric loss is combined and applied to the reconstruction phase. An optimization algorithm is then proposed to map the 3D face model to a 2D input head image [39]. The basic idea is to reconstruct a 3D model from the 2D input image to optimize efficient estimation. Monocular system and optical flow analysis [39] reconstruct a 3D head pose from 2D salient points.

The main essence of consumer technology (CT) is to map accurate 3D pose estimation from the given 2D image.

Experiences like user engagement, immersive audio, and user attentiveness can be supported seamlessly. Annotating the training or ground truth data is the essential part of the process and cannot be compromised on [40].

E. LOW QUALITY AND OPTIMIZATION

To deal with the limitations of low-quality images like noise, poor resolution, and occlusion [41], the Restricted Boltzmann Machine (RBM) model is proposed to map landmark locations to a 3D model by projection module and later predicts head pose with KL-divergence method and a gradient method. For optimizing the existing HPE methods, a key observation focuses on increasing the scale of bounding boxes in the captured image dataset to get more information from the input data and provide better results, as proposed in [41]. Some poor-quality images are shown in figure 4.

They also suggest that choosing the correct loss function can majorly impact the accuracy, as they have incorporated a pertained RESNET50 as the backbone of their CNN. Pose from Orthography and Scaling with Iterations (POSIT) and weighted POSIT (wPOSIT) algorithms are used to optimize existing state-of-the-art methodologies for estimating head pose [42].

III. PROPOSED METHODOLOGY

In this paper, head pose estimation is carried out by the sequential and ElasticNet regression models. The overview diagram is depicted in figure 5, where facial landmark determination is done initially. Then the CPAM model and sequential and ElasticNet regression models were implemented to calculate the desired outcomes for the head pose.

Figure 6 gives a detailed layout of the workflow of head pose estimation for attention span.

The sequential regression model is a linear stack model which is customized to give optimal results for HPE. ElasticNet model primarily focuses on bridging L1 and L2 losses and overcoming dependency of given sample data.

A. REGRESSION-BASED METHODS

Head pose estimation is a computer camera detecting a head's position in 3D spaces concerning the surrounding in an image or a video sequence. The space is relative to the camera [43]. The main goal of head pose estimation is to get three-dimensional Euler angles, including the pitch, roll, and yaw angles.

B. HEAD POSE ESTIMATION WITH CPAM

Coordinate Pair Angle Method (CPAM) shown in figure 7 uses 68 facial landmarks coordinates, which are coordinates points (x, y) on the given input image that outline the structure of the facial features given in Table 3. Following the landmark coordinates, we calculate the difference between the angles concerning the *x-axis* for each pair.

The functional diagram for CPAM and the proposed methodology followed for the implementation are shown in figure 7.

Hence, the proposed method consists of two steps:

- Extraction of 68 facial landmark coordinates from the given input face.
- Calculation of angular difference between each pair of coordinates taken w.r.t *x-axis* $(\theta = 0)$.

Consider a point A of coordinate A = (x1, y1) and a point B of coordinate B = (x2, y2) in equations 1 and 2; we calculate the angle between these two points in equation 3, which provides us with the base metric of HPE in CPAM w.r.t *x-axis*:

$$LENGTH = sqrt\ [(X1 - X2)^2 + (Y1 - Y2)^2] \tag{1}$$

$$ANGLE = cos^{-1}[(X2 - X1)/LENGTH]$$

$$If\ Y1 > Y2\ then, \tag{2}$$

$$ANGLE = 2PI - ANGLE\ (where\ 2PI\ is\ 360\ degrees) \tag{3}$$

Likewise, for every pair, the angular difference was calculated. As a result, a total number of (68*67/2) = 2278 features were obtained.

C. CPAM FOR HPE

The output array obtained through the CPAM is compared with the results of the regression model using DCNN and

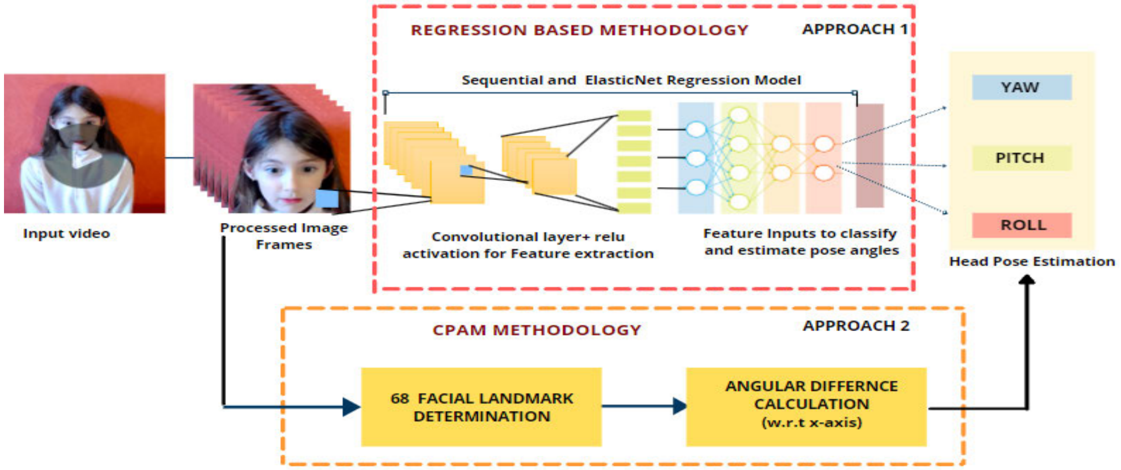

REGRESSION BASED METHODOLOGY
APPROACH 1
Sequential and ElasticNet Regression Model
Input video
Processed Image Frames
Convolutional layer+ relu activation for Feature extraction
Feature Inputs to classify and estimate pose angles
YAW
PITCH
ROLL
Head Pose Estimation
CPAM METHODOLOGY
APPROACH 2
68 FACIAL LANDMARK DETERMINATION
ANGULAR DIFFERNCE CALCULATION (w.r.t x-axis)

FIGURE 5. Overview of the proposed model.

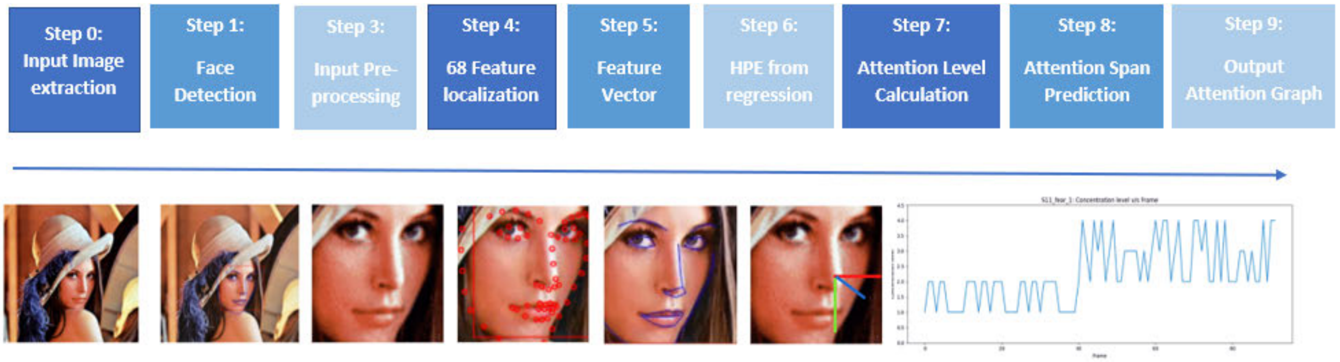

Step 0: Input Image extraction
Step 1: Face Detection
Step 3: Input Pre-processing
Step 4: 68 Feature localization
Step 5: Feature Vector
Step 6: HPE from regression
Step 7: Attention Level Calculation
Step 8: Attention Span Prediction
Step 9: Output Attention Graph

FIGURE 6. Stepwise implementation workflow for HPE to map attention span.

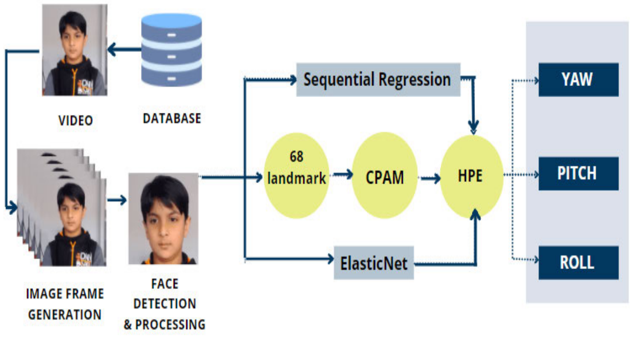


FIGURE 7. Functional diagram for CPAM and proposed methodology.

ElasticNet. Each of the features present in the array corresponds to an angular difference between the two coordinates.

The extracted vector from the input image gives the pose, i.e., *yaw, pitch, and roll.* Three different regression models are built for each axis.

Regression model prediction results for pitch, yaw, and roll are as follow:

Pitch in the range of [−65°, +65°]
Yaw in the range of [−90°, +90°]
Roll in the range of [−75°, +55°]

IV. DATASETS

To get a comparative result of existing HPE methodologies, the proposed model CPAM was trained on the 300W-LP dataset and tested with different datasets: 300W-LP dataset, AFLW2000 dataset, and NIMH-ChEFS dataset.

The 300W-LP dataset shown in figure 8 is an enlarged version of the 300W dataset. 61,225 samples are generated from profiling of 300W dataset and other sample databases: IBUG: 1,786 images, AFW: 5,207 images, LFPW: 16,556 images, HELEN: 37,676.

The AFLW2000 dataset shown in figure *9* offers images and corresponding annotations for about 2000 identities in AFLW (Annotated Facial Landmarks in the Wild). The 3D model is used for reannotating 68 3D landmarks. AFLW contains 25,000 RGB photos of faces which are taken from the social network Flickr. The sample images are highly diverse and comprise different lighting conditions, postures, occultations, attributes, and emotions. Since it is highly diverse, we consider this dataset for testing our model. Table 1 depicts the standard spread of pitch, yaw, and roll angles on the 300W-LP dataset, AFLW2000 dataset, and NIMH-ChEFS dataset.

FIGURE 8. Images selected from the 300W-LP dataset [44].

FIGURE 9. Images selected from the AFLW2000 dataset [45].

TABLE 1. The ranges for pitch, yaw, and roll value of the images from different datasets.

Dataset	NIMH-ChEFS	AFLW2000	300W-LP
Pitch	[-20^0, +20^0]	[-30^0, +30^0]	[-65^0, +65^0]
Yaw	[-30^0, +30^0]	[-45^0, +45^0]	[-90^0, +90^0]
Roll	[-25^0, +25^0]	[-20^0, +20^0]	[-75^0, +55^0]

National Institute of Mental Health Child Emotional Faces Picture Set (NIMH-ChEFS) dataset in figure 10 is a high-quality dataset of 482 photographs with colored images depicting emotions of a wide range of children. The gaze directions are either direct or averted.

FIGURE 10. Images selected from the NIMHS-ChEFS [46].

TABLE 2. Model parameters.

CNN Parameters	Quantity
Hidden Layer Count	3
Size of Each Layer	128
Learning Rate	0.0001
Epochs	30
Activation Function	Relu
Optimizer	Adam
Weight Decay	0.0001
Batch Size	65

TABLE 3. Location of landmark points of salient features.

Face Landmark Points	Landmark Locations
Jawline	1-17
Eyebrow	18-27
Nose	28-36
Eyes	37-78
Mouth	49-68

V. EXPERIMENTS AND RESULTS

A. MODEL SPECIFICATIONS

This methodology focuses on mapping 3- dimensional points from 2D image data. Sequential and ElasticNet model evaluations predict the experiment to form a comparative study. Given below in Table 2 are the parameters used to train the sequential model. 68 facial feature prediction is established through OpenCV in combination with the Dlib library.

The input is a face image given to the predictor; the output obtained is 68 facial landmark locations. Pi is the pair of Cartesian coordinates (xi, yi) where is the range of i is [1], [68]. The distribution of the coordinates over various face classes is depicted in Table 3. The output coordinates are the pixelated landmark locations. The obtained coordinated points are centered on a group of regression trees [47].

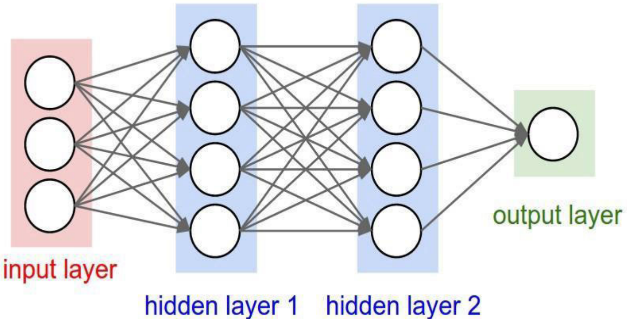


FIGURE 11. A basic feedforward neural network [53].

Traditionally, the mean human head model is used to map 2D facial key points to the 3D model of the targeted input image. In contradiction, [48] suggests that landmark detection is considered a fragile method. It implements a multi-loss convolutional neural network with both classification and regression models.

B. METHODOLOGY

1) COLOR/GRAYSCALE FIGURES

Supervised machine learning uses "labeled" datasets for training. It aims to map input variable x to the feature variable y through a function f(x) and predict the best fit on the given input data. Classification and regression are the two types of supervised learning [49]. One can make use of the regression supervised learning technique in this paper. The regression model usually works on continuous numeric data to produce a continuous output [50]. Here, the proposed approach leverages the sensitivity of the regression model to propose a best-fit prediction model and estimate unknown parameters for the three degrees of freedom for the pitch, yaw, and roll head pose angles.

2) DEEP NEURAL NETWORK REGRESSION

Neural Networks aim to mimic the human brain to understand the underlying association and patterns between a set of parameters [51], [52].

As shown in figure 11, a deep neural network uses layers to process the input data and provide an insightful output. It gives data a sophisticated touch by identifying underlying patterns. The nodes of the input layer are equal to the number of input features. A variable number of nodes are present in the hidden layers. Several layers in output are one if it is a regression task and several for classification tasks. The essential function of every node shown in figure 12 is to apply a weight to the incoming feature and add bias to it. Later, this is processed by an activation function like Relu, sigmoid or tan and returned to the next layer node. With the strong computational ability and robust design feature, the deep learning model keeps the features with maximum weight and contribution and discards the rest. It gives results based on a series of non-linear operations [33].

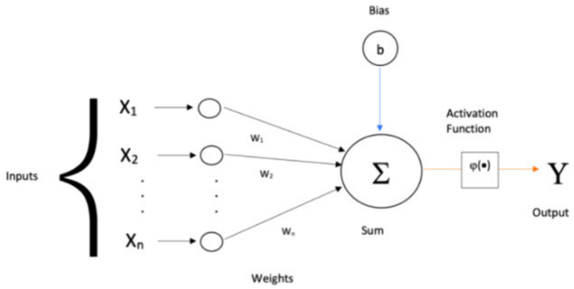


FIGURE 12. A single node in a neural network [54].

3) ELASTICNET REGRESSION

Elastic, known for incorporating the penalties from both L1 and L2 regularization [55], is one of the most used models for regression tasks. This model gives the most reliable output in highly correlated features and makes a precise selection through cross-validation [56].

$$\frac{1}{m}[\sum\nolimits_{l=1}^{m}(y^{(i)} - \mathrm{h}\left(x^{(i)}\right))^{2} + \lambda_1 \sum\nolimits_{j=1}^{n} w_j + \lambda_1 \sum\nolimits_{j=1}^{n} w^{2}{}_{j}] \tag{4}$$

In equation 4 w_j-weight for j$^{\text{th}}$ feature, n represents the number of features in the dataset, λ_1 and λ_2 are the regularization metrics for L1 and L2, respectively.

The terminology was derived from "the elements of statistical learning," a hyper-parameter "alpha" ($\boldsymbol{\alpha}$) is provided to assign the weight given to each of the L1 (ridge) term and L2 (Lasso) term [57]. ElasticNet allows us to use a million parameters for training. Its tuning feature selection for ridge and lasso when alpha is 0 or 1 gives this model an edge. For instance, let's take $\boldsymbol{\alpha}$ as 0, the penalty function is now L1 and if we take $\boldsymbol{\alpha}$ as 1, we get L2. Thus, optimizing the function will be dependent on the value of alpha [58].

For example, if $\boldsymbol{\alpha} = 0.5$, it would give a 50% contribution to the loss function of each penalty. This might shrink some coefficients and parameters, making them 0 and giving us a sparse result.

C. RESULTS

The proposed methodology was tested with the 300W-LP dataset, AFLW2000 dataset, and NIMH-ChEFS dataset to obtain a comparative study. The train test ratio of the above datasets was split into a 70:30 ratio described in the implementation section. Mean-Absolute-Error (*MAE*) is chosen as the evaluation metric for the proposed approach. The *MAE* measures the difference between the ground truth poses and the predicted value. It can be depicted by:

$$MAE = \frac{1}{n}\sum\nolimits_{j=1}^{n} |y - y^{\wedge}{}_{j}| \tag{5}$$

In equation 5, y is the true angular value poses, and yˆj is the predicted pose.

Table 4 and Table 5 show a comparative study of our proposed model, i.e., the CPAM-regression method evaluated on the 300W-LP Dataset, AFLW200, and NIMH-ChEFS

TABLE 4. The *MAE* on the AFLW2000 dataset.

Method	Yaw	Pitch	Roll	MAE
HopeNet	6.470	6.559	5.436	6.155
Hyperface	7.61	6.13	3.92	5.89
KEPLER	6.45	5.85	8.75	7.01
3DDFA	5.400	8.530	8.250	7.393
FAN	6.358	12.277	8.714	9.116
Dlib	23.153	13.633	10.545	15.777
QT-PYR	7.6	7.6	7.17	7.45
QuatNet	3.973	5.615	3.92	4.503
Proposed CPAM-DNNR	**1.479**	**1.804**	**1.809**	**1.697**
Proposed CPAM-ENR	**1.677**	**1.174**	**1.136**	**1.329**

TABLE 5. MAE on 300WW-LP dataset.

Method	Yaw	Pitch	Roll	Roll
CPAM-DNNR	2.504	1.189	2.350	2.014
CPAM-ENR	2.961	1.463	1.995	2.139

TABLE 6. MAE on NIMH CHEFS dataset.

Method	Yaw	Pitch	Roll	Roll
CPAM-DNNR	14.836	13.647	1.142	9.875
CPAM-ENR	15.247	14.265	1.323	10.278

dataset. The CPAM regression methods presented are Deep Neural Network Regression (CPAM- DNNR) and ElasticNet Regression (CPAM-ENR). Results show the *MAE* of head pose estimation obtained for HPE angles and total *MAE* error along with the three head pose angles. The least *MAE* value indicates the best model.

Table 6 displays the results of the NIMH-ChEFS Dataset. ChEFS dataset is mainly used to evaluate emotion detection; hence, this paper's HPE for NIMH-ChEFS is unique.

Figure 13 displays a graphical comparison with state-of-the-art methodologies. It shows a comparative result mapping of our model to the other state-of-the-art methodologies proposed on the AFLW2000 dataset. We observe that CPAM-DNNR and CPAM-ENR have exceptionally low MAE and hence are the best fit model compared to Hyperface, Multi-Loss Resnet, Quatnet, and more.

Figure 14 shows the estimated head pose of images from the NIMH-ChEFS dataset using CPAM-DNNR regression (left) and CPAM-ENR regression. Evaluations suggest the error to be less than 3°.

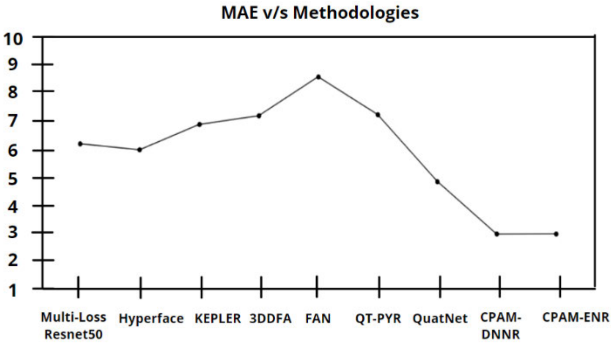


FIGURE 13. MAE on AFLW2000 dataset.

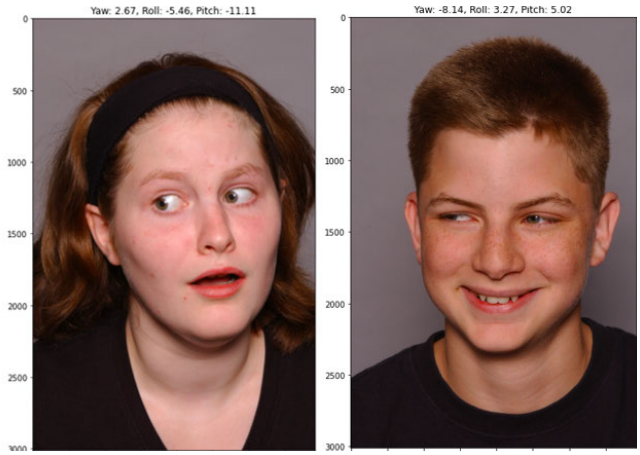

Yaw: 2.67, Roll: -5.46, Pitch: -11.11
Yaw: -8.14, Roll: 3.27, Pitch: 5.02
0
500
1000
1500
2000
2500
3000

FIGURE 14. Estimation of the head pose of images from the NIMH CHEFS dataset.

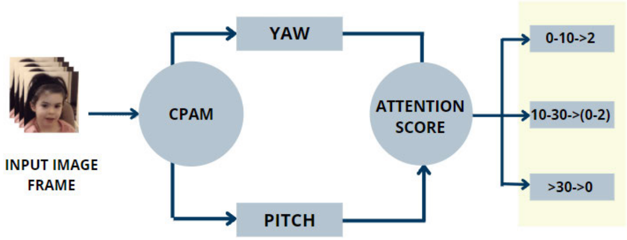
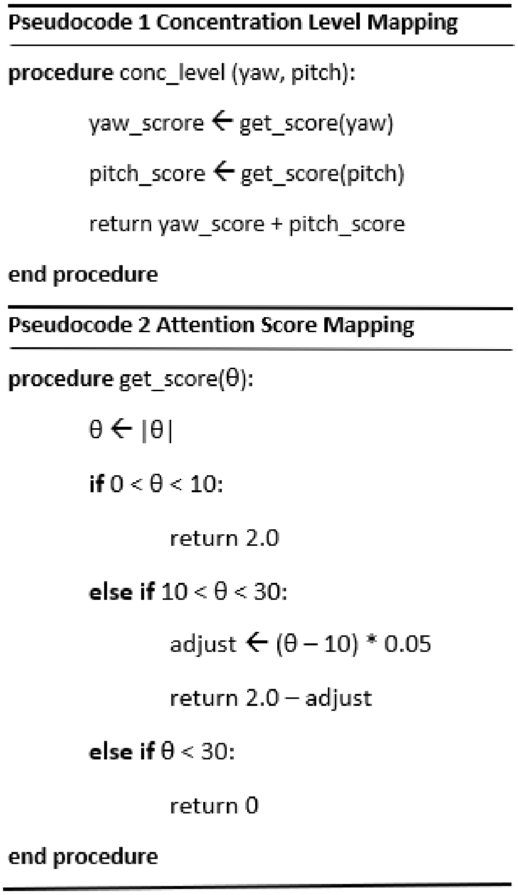


FIGURE 15. Proposed attention score prediction methodology.

D. DISCUSSION

The result comparison in Figure 13 shows that our model outperforms most of the optimal state-of-the-art methods, such as QuatNet [14], [59], Hyperface [60], Multi-Loss Resnet [61], [62] for HPE on the AFLW2000 dataset. NIMH-ChEFS dataset considers the emotions of children along with HPE. Our novel method has been thus tested upon a wide range of diverse real-world images with challenging limitations of occultation, low quality, and distorted data and yet produce a result with an MAE of 3° and less. The significance of this work extends the importance of considering facial key points, which become a crucial indicator of the head pose [43].



```
Pseudocode 1 Concentration Level Mapping
procedure conc_level (yaw, pitch):
        yaw_scrore ← get_score(yaw)
        pitch_score ← get_score(pitch)
        return yaw_score + pitch_score
end procedure
Pseudocode 2 Attention Score Mapping
procedure get_score(θ):
        θ ← |θ|
        if 0 < θ < 10:
                return 2.0
        else if 10 < θ < 30:
                adjust ← (θ – 10) * 0.05
                return 2.0 – adjust
        else if θ < 30:
                return 0
end procedure
```

FIGURE 16. Pseudocode to get attention span from the head pose.

VI. CALCULATION OF ATTENTION SPAN

A. ATTENTION MODULE METHODOLOGY

The concentration level of an individual at a given time while looking at a screen depends on their head orientation. Hence, concentration level and head orientation are correlated. While looking at a screen, *pitch* and *yaw* are the two main axial angles that play a vital role in predicting the concentration level. The attention span of an individual while looking at a screen can be determined with the average concentration level over a span of time. Hence, first, it is essential to formulate a methodology for getting the concentration level of an individual at a given frame with given *yaw* and *pitch* value. After getting the concentration level of each frame, its attention span was determined, as shown in figure 15.

B. ATTENTION SCORE IMPLEMENTATION

First, it is imperative to calculate the score factor of yaw angle and pitch angle, respectively, with the *get_score* function mentioned below. Then concentration level (Ci) at given frame i can be calculated with yaw and pitch value. The score variable ranges from [0, 2] for each given angle. Hence, the obtained concentration level at each frame was in a range [0, 4]. The proposed function to map the user's attention span from HPE can be seen in figure 16.

FIGURE 17. (a) Captured frame: S1_disgust_1, (b) Captured frame: S8_happy_3, (c) Captured frame: S11_fear_1.

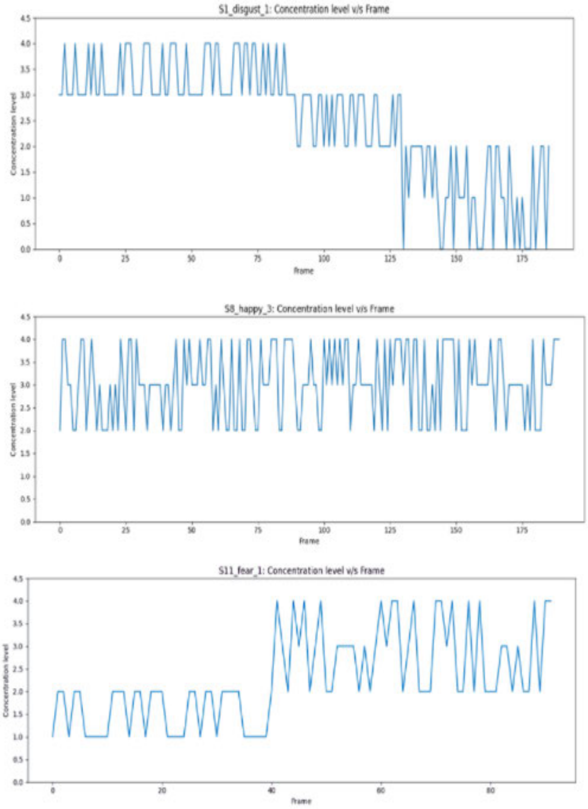


FIGURE 18. Attention graph of figure 17 (a, b, c) respectively.

After getting the concentration level for each of the 'n' frames, we can determine the attention span score with the help of equation 6:

$$Attention\ span\ score = \Sigma\ Ci/n \tag{6}$$

Next, the concentration level was classified based on attention score into four categories: no concentration, low concentration, medium concentration, and high concentration. Refer to Table 7 for mapping details.

C. VISUALIZATION OF RESULTS

To get the intuition of the above-discussed methodology, the LIRIS children dataset was used for visualization. LIRIS dataset contains 208 videos of 12 students between the age of [8], [12]. For visualization, first, the concentration level

TABLE 7. Attention SPAN score category.

Attention span score	Category
0 – 1.0	No concentration
1.1 – 2.0	Low concentration
2.1 – 3.0	Medium concentration
3.1 – 4.0	High concentration

for each frame (Ci) and attention span score was calculated with the help of these data. Then, the concentration level v/s frame graph was plotted to visualize the student's attention span.

Consequently, three videos were taken to visualize concentration in different scenarios while they showed different emotions. Figure 18 shows the attention graph of these children.

VII. CONCLUSION

Finding out attention span in online mode using head pose estimation is a challenging research problem. In this paper, ElasticNet and DCNN were separately used for HPE; the results obtained were then combined with the proposed methodology, i.e., Coordinate pair angle method (CPAM), to provide high precision results. The CPAM methodology exploits geometry to estimate head poses. When applied to specific standard datasets, the proposed methodology showed results on par with present state-of-the-art methodologies with 3° or less MAE in contrast to a standard of 6° MAE. Furthermore, the research formulated a method for predicting the attention span of an individual using head pose estimation, which can be used to visualize the concentration level for the entire period. This leads to defining this project's future scope targeting the eye gaze prediction module led by findings and estimation of roll and yaw orientations with image processing methods like the Viola-Jones algorithm. The authors intend to integrate the head pose estimation with gaze direction determination to propose an attention score. This research can be used in educational setups to estimate the understanding of the user; it can also be used in assistive driving environments to map the user's attention on the road. Further applications can be estimating concentration scores of users in medical setups for anxiety prediction, mental disturbance, and depression and mapping attention in the workplace environment. HPE and user emotion can help gauge engagement in social media advertising settings and improve sale strategies. Hence, this research is dynamic and can play a vital role in many applications.

AUTHOR CONTRIBUTIONS

Conceptualization, I.S., M.R., M.H., B.P.; methodology, I.S., M.R., M.H. B.P.; software, I.S. and M.H.; validation with field investigations, I.S. and M.R.; writing original draft preparation, I.S.; review and editing, B.P., C.-W.L., A.A.A.; supervision, M.R., M.H., and B.P.; funding acquisition, B.P., C.-W.L., and A.M.A. All authors have read and agreed to the published version of the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] T. Liu, J. Wang, B. Yang, and X. Wang, "NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom," *Neurocomputing*, vol. 436, pp. 210–220, May 2021, doi: 10.1016/J.NEUCOM.2020.12.090.

[2] P. Goldberg, Ö. Sämer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, "Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educ. Psychol. Rev.*, vol. 33, no. 1, pp. 27–49, Mar. 2021, doi: 10.1007/s10648-019-09514-z.

[3] S. Palmisano, R. S. Allison, and J. Kim, "Cybersickness in head-mounted displays is caused by differences in the user's virtual and physical head pose," *Frontiers Virtual Reality*, vol. 1, pp. 1–24, Nov. 2020, doi: 10.3389/frvir.2020.587698.

[4] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *Proc. Hum. Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 283–288, doi: 10.1109/ACII.2013.53.

[5] S. Wu, J. Liang, and J. Ho, "Head pose estimation and its application in TV viewers' behavior analysis," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2016, pp. 1–6, doi: 10.1109/CCECE.2016.7726649.

[6] Y. T. Hsieh and M. C. Yeh, "Head pose recommendation for taking good selfies," in *Proc. Workshop Multimodal Understand. Social, Affect. Subjective Attributes*, Oct. 2017, pp. 55–60, doi: 10.1145/3132515.3132518.

[7] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, "LipSync3D: Data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization," 2021, *arXiv:2106.04185*. [Online]. Available: http://arxiv.org/abs/2106.04185

[8] J. Meza, L. A. Romero, and A. G. Marrugo, "MarkerPose: Robust real-time planar target tracking for accurate stereo pose estimation," 2021, *arXiv:2105.00368*. [Online]. Available: http://arxiv.org/abs/2105.00368

[9] T. Hu, S. Jha, and C. Busso, "Robust driver head pose estimation in naturalistic conditions from point-cloud data," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1176–1182, doi: 10.1109/IV47402.2020.9304592.

[10] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognit.*, vol. 71, pp. 132–143, Nov. 2017, doi: 10.1016/j.patcog.2017.06.009.

[11] C. S. Indi, K. V. Pritham, V. Acharya, and K. Prakasha, "Detection of malpractice in E-exams by head pose and gaze estimation," *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 8, pp. 47–60, 2021, doi: 10.3991/ijet.v16i08.15995.

[12] M. Wang, D. Tao, and B. Huet, *Multimedia Modeling*, vol. 281. Cham, Switzerland: Springer, 2014.

[13] Z. Zhao, S. Xia, X. Xu, L. Zhang, H. Yan, Y. Xu, and Z. Zhang, "Driver distraction detection method based on continuous head pose estimation," *Comput. Intell. Neurosci.*, vol. 2020, Nov. 2020, Art. no. 9606908, doi: 10.1155/2020/9606908.

[14] R. Valle, J. M. Buenaposada, and L. Baumela, "Multi-task head pose estimation in-the-wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 8, pp. 2874–2881, Aug. 2020, doi: 10.1109/TPAMI.2020.3046323.

[15] G. Sang and S. Xuan, "Learning toward practical head pose estimation," *Opt. Eng.*, vol. 56, no. 8, 2017, Art. no. 083104, doi: 10.1117/1.oe.56.8.083104.

[16] P. Li, Y. Li, and L. Tan, "Transfer useful knowledge for headpose estimation from low resolution images," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 9395–9408, Aug. 2016, doi: 10.1007/s11042-016-3297-2.

[17] P. Sreekanth, U. Kulkarni, S. Shetty, and M. S. M., "Head pose estimation using transfer learning," in *Proc. Int. Conf. Recent Trends Advance Comput. (ICRTAC)*, Sep. 2018, pp. 73–79, doi: 10.1109/ICRTAC.2018.8679209.

[18] J. Han and Y. Liu, "Head posture detection with embedded attention model," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 782, no. 3, 2020, Art. no. 032003, doi: 10.1088/1757-899X/782/3/032003.

[19] Y. Yu, K. A. F. Mora, and J.-M. Odobez, ''HeadFusion: 360 head pose tracking combining 3D morphable model and 3D reconstruction,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2653–2667, Nov. 2018, doi: 10.1109/TPAMI.2018.2841403.

[20] H. Yuan, M. Li, J. Hou, and J. Xiao, ''Single image-based head pose estimation with spherical parametrization and 3D morphing,'' *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107316, doi: 10.1016/j.patcog.2020.107316.

[21] S. Basak, P. Corcoran, F. Khan, R. Mcdonnell, and M. Schukat, ''Learning 3D head pose from synthetic data: A semi-supervised approach,'' *IEEE Access*, vol. 9, pp. 37557–37573, 2021, doi: 10.1109/ACCESS.2021.3063884.

[22] E. Murphy-Chutorian and M. M. Trivedi, ''Head pose estimation in computer vision: A survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009, doi: 10.1109/TPAMI.2008.106.

[23] O. Younis, W. Al-Nuaimy, M. H., and F. Rowe, ''A hazard detection and tracking system for people with peripheral vision loss using smart glasses and augmented reality,'' *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 1–9, 2019, doi: 10.14569/ijacsa.2019.0100201.

[24] A. F. Abate, P. Barra, C. Pero, and M. Tucci, ''Head pose estimation by regression algorithm,'' *Pattern Recognit. Lett.*, vol. 140, pp. 179–185, Dec. 2020, doi: 10.1016/j.patrec.2020.10.003.

[25] A. F. Abate, P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, ''Near real-time three axis head pose estimation without training,'' *IEEE Access*, vol. 7, pp. 64256–64265, 2019, doi: 10.1109/ACCESS.2019.2917451.

[26] K. Khan, J. Ali, K. Ahmad, A. Gul, G. Sarwar, S. Khan, Q. T. H. Ta, T.-S. Chung, and M. Attique, ''3D head pose estimation through facial features and deep convolutional neural networks,'' *Comput., Mater. Continua*, vol. 66, no. 2, pp. 1757–1770, 2021, doi: 10.32604/cmc.2020.013590.

[27] J. Xia, L. Cao, G. Zhang, and J. Liao, ''Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks,'' *IEEE Access*, vol. 7, pp. 48470–48483, 2019, doi: 10.1109/ACCESS.2019.2909327.

[28] A. Gupta, K. Thakkar, V. Gandhi, and P. J. Narayanan, ''Nose, eyes and ears: Head pose estimation by locating facial keypoints,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1977–1981, doi: 10.1109/ICASSP.2019.8683503.

[29] R. Berral-Soler, F. J. Madrid-Cuevas, R. Muñoz-Salinas, and M. J. Marín-Jiménez, ''RealHePoNet: A robust single-stage ConvNet for head pose estimation in the wild,'' *Neural Comput. Appl.*, vol. 33, no. 13, pp. 7673–7689, Jul. 2021, doi: 10.1007/s00521-020-05511-4.

[30] I. Aljarrah, A. Idries, and M. A. Atahat. (Sep. 2012). *A New Spatial Compression Algorithm for Colored and Gray-Scale Images*. [Online]. Available: https://www.researchgate.net/publication/232815017_A_New_Spatial_Compression_Algorithm_for_Colored_and_Gray-Scale_Images

[31] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, ''Head pose estimation on low resolution images,'' in *Multimodal Technologies for Perception of Humans* (Lecture Notes in Computer Science), vol. 4122. Cham, Switzerland: Springer, 2007, pp. 270–280, 2007, doi: 10.1007/978-3-540-69568-4_24.

[32] X. Xu and I. A. Kakadiaris, ''Joint head pose estimation and face alignment framework using global and local CNN features,'' in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 642–649, doi: 10.1109/FG.2017.81.

[33] J. M. D. Barros, F. Garcia, B. Mirbach, K. Varanasi, and D. Stricker, ''Combined framework for real-time head pose estimation using facial landmark detection and salient feature tracking,'' in *Proc. 13th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2018, pp. 123–133, doi: 10.5220/0006628701230133.

[34] J. Wang, J. Zhang, C. Luo, and F. Chen, ''Joint head pose and facial landmark regression from depth images,'' *Comput. Vis. Media*, vol. 3, no. 3, pp. 229–241, Sep. 2017, doi: 10.1007/s41095-017-0082-8.

[35] H. Liu, H. Nie, Z. Zhang, and Y.-F. Li, ''Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction,'' *Neurocomputing*, vol. 433, pp. 310–322, Apr. 2021, doi: 10.1016/j.neucom.2020.09.068.

[36] A. Behera, Z. Wharton, P. Hewage, and S. Kumar, ''Rotation axis focused attention network (RAFA-Net) for estimating head pose,'' in *Proc. Asian Conf. Comput. Vis.* (Lecture Notes in Computer Science), vol. 12626. Cham, Switzerland: Springer, 2021, pp. 223–240, doi: 10.1007/978-3-030-69541-5_14.

[37] Z. Cao, Z. Chu, D. Liu, and Y. Chen, ''A vector-based representation to enhance head pose estimation,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1187–1196, doi: 10.1109/wacv48630.2021.00123.

[38] C. Li, F. Zhong, Q. Zhang, and X. Qin, ''Accurate and fast 3D head pose estimation with noisy RGBD images,'' *Multimedia Tools Appl.*, vol. 77, no. 12, pp. 14605–14624, Jun. 2018, doi: 10.1007/s11042-017-5050-x.

[39] L. Liu, Z. Ke, J. Huo, and J. Chen, ''Head pose estimation through keypoints matching between reconstructed 3D face model and 2D image,'' *Sensors*, vol. 21, no. 5, pp. 1–24, 2021, doi: 10.3390/s21051841.

[40] S. Basak, F. Khan, R. McDonnell, and M. Schukat, ''Learning accurate head pose for consumer technology from 3D synthetic data,'' in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2021, pp. 1–6, doi: 10.1109/ICCE50685.2021.9427768.

[41] M. Shao, Z. Sun, M. Ozay, and T. Okatani, ''Improving head pose estimation with a combined loss and bounding box margin adjustment,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–4, doi: 10.1109/FG.2019.8756605.

[42] M. Ariz, A. Villanueva, and R. Cabeza, ''Robust and accurate 2D-tracking-based 3D positioning method: Application to head pose estimation,'' *Comput. Vis. Image Understand.*, vol. 180, pp. 13–22, Mar. 2019, doi: 10.1016/j.cviu.2019.01.002.

[43] J. Li, J. Wang, and F. Ullah, ''An end-to-end task-simplified and anchor-guided deep learning framework for image-based head pose estimation,'' *IEEE Access*, vol. 8, pp. 42458–42468, 2020, doi: 10.1109/ACCESS.2020.2977346.

[44] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, ''Feature reconstruction disentangling for pose-invariant face recognition supplementary material,'' *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1623–1632.

[45] H. Yi, C. Li, Q. Cao, X. Shen, S. Li, G. Wang, and Y.-W. Tai, ''MMFace: A multi-metric regression network for unconstrained face reconstruction,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7655–7664, doi: 10.1109/CVPR.2019.00785.

[46] H. L. Egger, D. S. Pine, E. Nelson, E. Leibenluft, M. Ernst, K. E. Towbin, and A. Angold, ''The NIMH child emotional faces picture set (NIMH-ChEFS): A new set of children's facial emotion stimuli,'' *Int. J. Methods Psychiatric Res.*, vol. 20, no. 3, pp. 145–156, Sep. 2011, doi: 10.1002/mpr.343.

[47] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, ''Web-shaped model for head pose estimation: An approach for best exemplar selection,'' *IEEE Trans. Image Process.*, vol. 29, pp. 5457–5468, 2020, doi: 10.1109/TIP.2020.2984373.

[48] N. Ruiz, E. Chong, and J. M. Rehg, ''Fine-grained head pose estimation without keypoints,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2155–2164, doi: 10.1109/CVPRW.2018.00281.

[49] V. Nasteski, ''An overview of the supervised machine learning methods,'' *Horizons B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[50] R. Choudhary and H. K. Gianey, ''Comprehensive review on supervised machine learning algorithms,'' in *Proc. Int. Conf. Mach. Learn. Data Sci. (MLDS)*, Dec. 2017, pp. 38–43, doi: 10.1109/MLDS.2017.11.

[51] C. K. Goh, Y. Liu, and A. W. K. Kong, ''A constrained deep neural network for ordinal regression,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 831–839, doi: 10.1109/CVPR.2018.00093.

[52] J. Du and Y. Xu, ''Hierarchical deep neural network for multivariate regression,'' *Pattern Recognit.*, vol. 63, pp. 149–157, Mar. 2017, doi: 10.1016/j.patcog.2016.10.003.

[53] S. B. Maind and P. Wankar, ''Research paper on basic of artificial neural network,'' *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. 1, pp. 96–100, 2014.

[54] A. D. Dongare, R. R. Kharde, and A. D. Kachare, ''Introduction to artificial neural network (ANN) methods,'' *Int. J. Eng. Innov. Technol.*, vol. 2, no. 1, pp. 189–194, 2012. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1082.1323&rep=rep1&type=pdf

[55] M. Talpade, D. Lynch, B. Lattimore, and A. Graham, ''The juvenile and adolescent substance abuse prevention program: An evaluation,'' *Int. J. Behav. Consultation Therapy*, vol. 4, no. 4, pp. 304–310, 2008, doi: 10.1037/h0100860.

[56] R. Alhamzawi and H. T. M. Ali, ''The Bayesian elastic net regression,'' *Commun. Statist.-Simul. Comput.*, vol. 47, no. 4, pp. 1168–1178, Apr. 2018, doi: 10.1080/03610918.2017.1307399.

[57] E. Pamukcu, ''Choosing the optimal hybrid covariance estimators in adaptive elastic net regression models using information complexity,'' *J. Stat. Comput. Simul.*, vol. 89, no. 16, pp. 2983–2996, 2019, doi: 10.1080/00949655.2019.1647431.

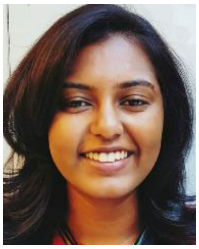
[58] H. Zou and T. Hastie, "Addendum: Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 5, p. 768, 2005, doi: 10.1111/j.1467-9868.2005.00527.x.

[59] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019, doi: 10.1109/TMM.2018.2866770.

[60] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[61] S. Wang, J. Li, P. Yang, T. Gao, A. R. Bowers, and G. Luo, "Towards wide range tracking of head scanning movement in driving," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 34, no. 13, Apr. 2020, Art. no. 2050033, doi: 10.1142/S0218001420500330.

[62] Y. Zhang, K. Fu, J. Wang, and P. Cheng, "Learning from discrete Gaussian label distribution and spatial channel-aware residual attention for head pose estimation," *Neurocomputing*, vol. 407, pp. 259–269, Sep. 2020, doi: 10.1016/J.NEUCOM.2020.05.010.



TRIPTI SINGH is currently pursuing the B.Tech. degree from the Information Technology Department, Symbiosis Institute of Technology, Symbiosis International University, Pune. She is an Intern with the Symbiosis Centre of Applied Artificial Intelligence (SCAAI). Her research interests include artificial intelligence, machine learning, deep learning, and computer vision.

MOHAN MOHADIKAR received the degree from the Computer Science Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. His research interests include computer vision, deep learning, machine learning, and artificial intelligence.

SHILPA GITE received the Ph.D. degree in deep learning for assistive driving in semi-autonomous vehicles from Symbiosis International (Deemed University), Pune, India, in 2019. She is currently working as an Associate Professor with the Computer Science Department, Symbiosis Institute of Technology, Pune. She is also working as an Associate Faculty with the Symbiosis Centre of Applied AI (SCAAI). She has around 13 years of teaching experience. She is also guiding Ph.D. students in biomedical imaging, self-driving cars, and natural language processing areas. She has published more than 50 research papers in international journals and 20 Scopus indexed international conferences. Her research interests include deep learning, machine learning, medical imaging, and computer vision. She was also a recipient of Best Paper Award at 11th IEMERA Conference held virtually with the Imperial College, London, in October 2020.

SHRUTI PATIL received the M.Tech. degree in computer science and the Ph.D. degree in the domain of data privacy from Pune University. She has been an Industry Professional in the past, currently associated with the Symbiosis Institute of Technology, as a Professor and with SCAAI, as a Research Associate, Maharashtra, Pune. She has three years of industry experience and ten years of academic experience. She has expertise in applying innovative technology solutions to real world problems. She is currently working in the application domains of healthcare, sentiment analysis, emotion detection and machine simulation via which she is also guiding several U.G., P.G., and Ph.D. students as a domain expert. She has published more than 30 research articles in reputed international conferences and Scopus/ web of science indexed journals, books with more than 100 citations. Her research interests include applied artificial intelligence, natural language processing, acoustic AI, adversarial machine learning, data privacy, digital twin applications, GANS, and multimodal data analysis.

BISWAJEET PRADHAN received the Habilitation degree in remote sensing from the Dresden University of Technology, Germany, in 2011. He is currently the Director of the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT. He is also a Distinguished Professor with the University of Technology Sydney. He is also an Internationally Established Scientist in geospatial information systems (GIS), remote sensing and image processing, complex modeling/geo-computing, machine learning, soft-computing applications, natural hazards, and environmental modeling. From 2015 to 2021, he worked as the Ambassador Scientist for the Alexander Humboldt Foundation, Germany. More than 650 articles, more than 550 have been published in science citation index (SCI/SCIE) technical journals. In addition, he has authored eight books and 13 book chapters. He was a recipient of the Alexander von Humboldt Fellowship from Germany. He has been received 55 awards in recognition of his excellence in teaching, service, and research, since 2006. He was also a recipient of the Alexander von Humboldt Research Fellowship from Germany. From 2016 to 2020, he was listed as the Highly Cited Researcher by Clarivate Analytics Report as one of the world’’s most influential mind. In 2018-2020, he was awarded as the World Class Professor by the Ministry of Research, Technology and Higher Education, Indonesia. He is also an Associate Editor and an Editorial Member of more than eight ISI journals. He has widely traveled abroad, visiting more than 52 countries to present his research findings.

ABDULLAH ALAMRI received the B.S. degree in geology from King Saud University, in 1981, the M.Sc. degree in applied geophysics from the University of South Florida, Tampa, in 1985, the Ph.D. degree in earthquake seismology from the University of Minnesota, USA, in 1990, and the M.S. degree. He is currently a Professor in earthquake seismology, the Director of the Seismic Studies Center, King Saud University (KSU). He is also the President of the Saudi Society of Geosciences and the Editor-in-Chief of the Arabian Journal of Geosciences (AJGS). He is a member of the Seismological Society of America, American Geophysical Union, European Assessment for Environmental and Engineering Geophysics, Earthquakes Mitigation in the Eastern Mediterranean Region, National Commission for Assessment and Mitigation of Earthquake Hazards in Saudi Arabia, Mitigation of Natural Hazards Com at Civil Defense. His research interests include the area of crustal structures and seismic micro zoning of the Arabian Peninsula. His recent projects involve also applications of EM and MT in deep groundwater exploration of Empty Quarter and geothermal prospecting of Volcanic Harrats in the Arabian shield. He has published more than 150 research papers, achieved more than 45 research projects and authored several books and technical reports. He is the principal and a co-investigator in several national and international projects (KSU, KACST, NPST, IRIS, CTBTO, U.S. Air force, NSF, UCSD, LLNL, OSU, PSU, and Max Planck). He has also chaired and co-chaired several SSG, GSF, and RELEMR workshops and forums in the Middle East. He obtained several worldwide prizes and awards for his scientific excellence and innovation.