BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **24**, No 2
Sofia • 2024 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2024-0018

A Novel Deep Transfer Learning-Based Approach for Face Pose Estimation

Mayank Kumar Rusia[1], *Dushyant Kumar Singh*[1], *Mohd. Aquib Ansari*[2]

[1]*Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, Uttar Pradesh, India*
[2]*SCSET, Bennett University, Greater Noida, U.P., India*
E-mails: mayank.qip18@mnnit.ac.in dushyant@mnnit.ac.in mansari.aquib@gmail.com

***Abstract*:** *An efficient face recognition system is essential for security and authentication-based applications. However, real-time face recognition systems have a few significant concerns, including face pose orientations. In the last decade, numerous solutions have been introduced to estimate distinct face pose orientations. Nevertheless, these solutions must be adequately addressed for the three main face pose orientations: Yaw, Pitch, and Roll. This paper proposed a novel deep transfer learning-based multitasking approach for solving three integrated tasks, i.e., face detection, landmarks detection, and face pose estimation. The face pose variation vulnerability has been intensely investigated here underlying three modules: image preprocessing, feature extraction module through deep transfer learning, and regression module for estimating the face poses. The experiments are performed on the well-known benchmark dataset Annotated Faces in the Wild* (*AFW*)*. We evaluate the outcomes of the experiments to reveal that our proposed approach is superior to other recently available solutions.*

***Keywords*:** *Face alignment, Biometrics, Face recognition, Image processing, Landmark detection, Deep convolutional neural network.*

1. Introduction

The global demand for face biometric-based identification systems is increasing significantly because of their non-intrusiveness nature, uniqueness, accessibility, and other related factors [1, 2]. Face recognition systems can easily be implemented for any computer vision application, such as color-texture-based face tracking [3, 4], border surveillance, face verification, human detection [5], and more. The process of locating the human face region in the entire image is termed face detection, which is the primary task of face recognition. Obtaining a frontal pose of the face before the verification/acquisition device is imperative for achieving accurate results. However, this often differs from the scenario in real-time systems, leading to the challenge of different face pose orientations. Current advancements in the field address the issue

of face pose variations by considering the angular displacement or rotation of face alignment in three dimensions, typically represented by Yaw, Pitch, and Roll angles, relative to the frontal position, as shown in Fig. 1.

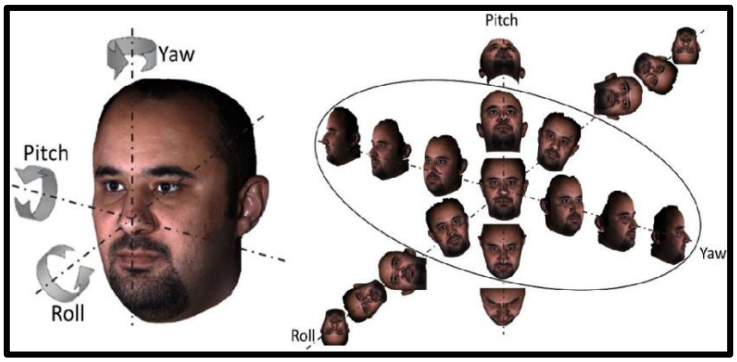


Fig. 1. Possible face pose orientations

Yaw, pitch, and roll are terms commonly used to describe the orientation and position of the head in three-dimensional space. These terms can be defined as follows:

- Yaw refers to the rotation around the vertical axis of an object. In the context of head position, it represents a left or right rotation of the face when viewed from the front. For instance, turning your head to the left or right without tilting it up or down is considered a yaw position. The yaw position can be identified and tracked from the center of gravity (i.e., Nose center) point, which is directly perpendicular to the right and left displacement of the head.
- Pitch position can also be termed the transverse position, which means the rotation of the face around its horizontal axis. More specifically, it represents the head's lateral movement (i.e., up or down tilt) from the center of gravity. For instance, if you nod your head up and down without turning it to the side.
- Roll corresponds to the rotation of the face around the forward-facing axis. When it comes to face position, it represents the tilting of the face from one side to the other. Roll can be described as a longitudinal axis concerning the point of center gravity (i.e., Nose center) and directed in parallel to the line of reference. For illustration, if you tilt your head to the left or right, you alter the roll position.

Face pose variation is the most common threat to face recognition systems, where a fraudster can bypass the verification device or benefit from the doubt by showing their profile or tiled face.

1.1. Motivation

The following motivational factors have inspired us to do this research work.

- A particular face posture reflects a person's gaze, psychology, and intention, which can help to analyze that person's behavior [6].
- Various face pose orientations (right, left, up, down, and round directions) are also prominent ways to represent the liveness of the face [7].

- Identifying the correct face, mainly when face images contain extreme poses, is still challenging [8].
- Convolutional neural network-based architecture has presented an excellent performance recently for image-based classification and regression problems, such as face recognition, object detection, activity recognition, image and video classification, and more.

1.2. Problem statement

An intelligent face-recognition system must be capable of classifying the face and non-face images and estimating the facial landmarks and distinct face poses. More specifically, the verification system should be powered enough to detect the face region from the input frame (if the face exists). Additionally, the system should be able to estimate the various face poses ranging from –30° to +30° (i.e., –30°, –15°, 0°, +15°, +30°) that comprise Yaw, Pitch, and roll direction.

The verification system needs to examine whether the claimed face identity is the same as intended (i.e., same as stored in the database) through the two separate checks, first is the classification, and second is the regression. Based on automatic prediction and generated loss function, the estimation of different pose orientations of the claimed person is evaluated. Here, the hypothesis (i.e., presumption) is that if the input image has a face, then the face image must have a frontal, right, left, top, and down profile face. Here, the hypothesis (i.e., presumption) is that if the input image has a face, then the face image must have a frontal, right, left, top, and down profile face.

1.3. Contribution

We aimed to deploy a deep neural network-based multitasking approach that involves three proposed models, i.e., face detection, fiducial point detection, and face pose estimation. All these models have been trained on fused deep transfer learning-based architectures, consisting of InceptionV3, VGG16, ResNet50, and MobileNetV2. The discrimination between the frontal and non-frontal faces (profile images) is identified through the face detection model, whereas fiducial points for the face attributes have been detected through sixty-eight landmarks-based shape predictors utilized from the Dlib library. The face pose estimation has been analyzed through a regression model. Our major contribution can be summarized below:

- A Dlib shape predictor library detects the human's face region from the input image based on 68 landmark points.
- To avoid the useless training of numerous parameters and features, we customized the Dlib shape predictor with a reduced number of the identified 68 landmarks to only six essential points (i.e., nose, chin, left eye and right eye corner points, and corner points of mouth) which are enough to detect and estimate the face pose variations.
- This work utilizes the four most prominent deep transfer learning-based models with optimized hyperparameters tuning for the training and validation tasks performed on a well-known AFW benchmark dataset.

- We investigated the performance of all the proposed models and compared them with other existing methods to demonstrate the effectiveness of challenging unconstrained datasets.

1.4. Organization

This paper is organized into five sections. Section 1 delineates a rationale behind this work with the valid motivation to develop an advanced face recognition system invariant to face pose variations. Section 2 discusses the related work supporting the study of face detection, various face poses orientations, multitasking, and deep neural networks-based state-of-the-art techniques. Section 3 illustrates the proposed methodology for data pre-processing, feature engineering, feature selection, and multi-tasking approach for classification and regression. Section 4 demonstrates the comparative analysis of achieved results with other existing methods. Section 5 summarizes the work followed by the future scope in the last of the article.

2. Related work

Extensive research has been done for face detection and different face pose identification. This section briefly discusses a brief of the work done in the domain of face detection, multi-tasking, deep neural network-based techniques, and face pose estimation. In the recent past, face detection research has attracted the attention of researchers to a large extent. The solution for various challenges associated with face detection and recognition systems, such as face pose variations, facial expressions, and more has been introduced in the last decade.

Wu, Zhang and Tian [9] introduced a multi-tasking cascaded framework consisting of two convolutional neural networks for solving two integrated tasks: face detection and pose estimation. The face detection and pose estimation tasks have been evaluated on FDDB and AFW benchmark datasets, respectively. The experimental results show that the multi-tasking CNN method is superior to other face detection and pose estimation methods. Ranjan, Patel and Chellappa [10] presented an algorithm for four simultaneous tasks (i.e., face detection, landmark detection, pose estimation, and gender recognition), implemented using a deep convolutional neural network called Hyperface. The two variants of Hyperface are ResNet-101 (performance improvement) and Fast-Hyperface (speed improvement). The HF-ResNet represents a minimum Normalized Mean Error (NME) of 2.71% for [0,30] face alignment cases with an overall mean of 2.93% and a Standard deviation of 0.25% on the AFLW dataset. The minimum NME for landmark detection is 8.18% for HF-ResNet on the IBUG dataset. While HF-Face represents less pose estimation error of 97.7% for the AFW dataset with $\pm 15°$ error tolerance. HF-ResNet shows a better accuracy for gender recognition for both the celebs and LFWA datasets.

Zhang et al. [11] proposed a deep cascaded multi-task framework that incorporates the inherent correlation between multitasking to predict the face and its location through landmarks in a coarse-to-fine manner. The challenging FDDB and WIDER FACE benchmark datasets are utilized for face detection, while the AFLW

benchmark datasets are used for face alignment. An et al. [12] proposed a novel face alignment method named Adaptive Pose Alignment (APA) to estimate the face pose variation. This method aims to learn the alignment template, including two individual tasks first is to reduce the intra-class difference, and the second is to reduce the noise during traditional alignment. The IJB-A, IJB-C, and CPLFW datasets are used for experimentation purposes. The proposed method has achieved an accuracy of 99.80% on the LFW dataset and 92.95% on the CPLFW dataset. He et al. [13] proposed a novel Deformable Face Net (DFN) method named to analyze the pose variations for efficient face recognition. Here, the network has learned alignment-oriented and identity-preserving features of the faces. To evaluate the model's performance, two loss functions are considered: Identity Consistency Loss (ICL) and the Pose Triplet Loss (PTL). The experiments reveal that the proposed method outperforms the other recent methods, especially on extreme pose datasets.

Han et al. [14] introduced a novel face pose estimation method that integrates VGG-Face and multi-scale Curvelet representation. The VGG-Face representation utilized a CNN model as a backbone with additional transfer learning. The Mean Absolute Error (MAE) of 0.33° and 0.23° has been achieved for the Yaw and pitch angle on the CAS-PEAL pose database. Masi et al. [15] proposed a novel method to tackle the problem of extreme face pose variations instead of using a single model to learn pose invariance through the massive amount of data or to normalize the face images to a single frontal pose. Instead, the deep CNN-based proposed Pose-Aware Models (PAM) model with 3D rendering synthesizes distinct face poses. Comparative evaluations of this technique for both IJB-A and PIPA datasets are performed. Fard, Abdollahi and Mahoor [16] proposed an active shape model-based method (i.e., ASMNet) to locate the target object (i.e., face), and lightweight CNN (i.e., MobileNetV2) is utilized to align the face and estimate the pose of the face onto an image. The results of the experiments reveal that ASMNet achieves an acceptable performance on a challenging dataset with a significantly smaller number of parameters.

Yin and Liu [17] introduced Multi-Tasking Learning (MTL) with a CNN model to identify the face from a given image. using a dynamic weighting scheme, this method also provides solutions for other side tasks related to facial recognition, such as pose, illumination, and expression estimation. A new CNN model for the pose-specific feature learning and energy-based weight analysis method is also proposed here. The multi-PIE dataset performs the face detection and pose variation experiments with a comparative analysis of LFW, CFP, and IJB-A datasets results.

The state-of-the-art techniques discussed above convey two solutions for face pose variation problems. One is a deep learning-based single task, and the other is a deep learning-based multi-task. The multitask solution approach is always superior as this method can easily tackle face detection and generate loss functions for detecting landmarks and estimating the face pose. Therefore, this work proposed a deep neural network-based multi-tasking approach to detect the face and landmarks of the face and assess the distinct face poses. The next section discusses a systematic methodology.

3. Proposed methodology

The recent development in deep neural network-based approaches represents outstanding results for classification and regression problems. We proposed a novel deep transfer learning-based approach to classify the input frame into frontal and non-frontal faces along with an estimation of the face poses in three different orientations, i.e., Yaw, Pitch, and Roll position. The complete process of the proposed model consists of three processes: image preprocessing, feature extraction through deep transfer learning, and regression-based estimation of different face poses based on regression. The general framework of the proposed methodology is shown in Fig. 1.

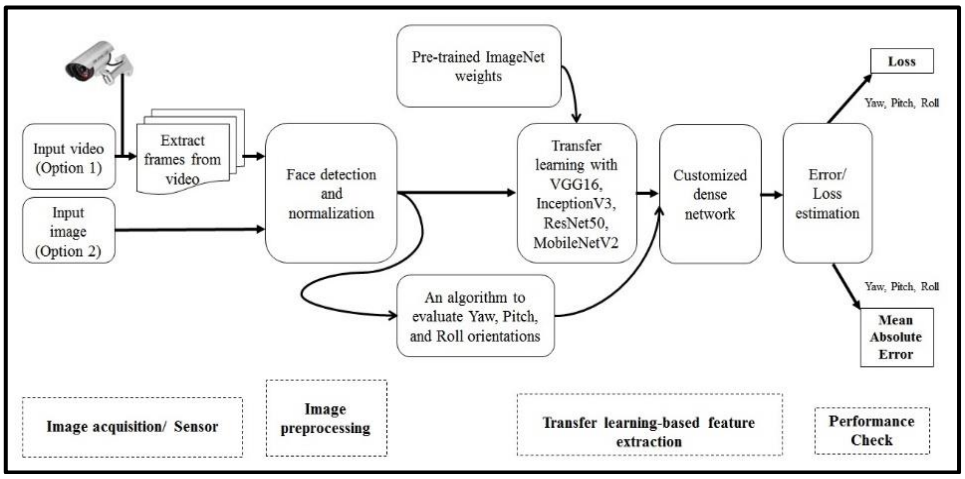


Fig. 2. A general framework of the proposed model

The next sub-section illustrates a systematic process of image preprocessing after capturing an input frame.

3.1. Image preprocessing module

Image preprocessing is a set of functions and operations performed on images before the feature extraction process to enhance the quality of the input images or video frames. Our proposed approach involves various image processing techniques, such as grayscale conversion, resizing, face detection, segmentation, normalization, and augmentation. Grayscale is a single intensity-based representation used to discriminate the low-level information. Our proposed model deals with grayscale channels as it does not require any color information. Therefore, we convert the RGB color channel input data to the grayscale channel for fast computing and efficient results. Resizing is a process of changing the dimension of an object while preserving the aspect ratio. The images that are processed with deep learning methods, especially transfer learning models must have a predefined input size as per the requirement of the algorithm. The other operations on image processing are described as follows:

- Face detection is a primary task in any face biometric-based problems. We deployed a Dlib shape predictor consisting of 68 facial landmarks, also termed fiducial points for performing two specific tasks, such as face detection and fiducial point detection. Dlib is a pre-trained model that extracts the facial regions from an input image and subsequently locates the 68 fiducial points on the facial attributes, including eyes, nose, jawline, eyebrows, and inner, and outer mouth. These distinctive points have proven effective in various applications such as aligning faces, estimating head poses, swapping faces, detecting blinks, and more. This model returns a list of rectangles representing the bounding boxes of the detected faces.
- Segmentation is a process of segregating a specific region (the face) and discarding other useless information about the image. Once the face points are identified by a Dlib shape predictor a trace covering all fiducial points generates a bounding box, which is finally cropped (or segmented). This function can be used to eliminate background or unwanted objects from the region of interest.
- Normalization is a process to scale the corresponding pixel values to a standard range of allowed intensity values. We normalize our data as per the architectures proposed for feature extraction.
- Augmentation is a process to enlarge the capacity of the available data to provide more samples to the feature extraction module for efficient training and subsequently learning the patterns. Augmentation is mostly utilized for the cases where size of the datasets is small. However, the augmentation process applies various random transformations to generate multiple variants from a single image, available only for training not for testing. The random transformation for the augmentation process includes rotation, translation, scaling, flipping, and shearing.

3.2. Feature extraction module

Estimating the various face pose orientations is a tough task. However, our proposed algorithm provides a simplified approach to estimate the yaw, pitch, and roll positions based on the objective of this work. We deployed transfer learning-based models consisting of fine-tuned VGG16, InceptionV3, ResNet50, and MobileNetV2 architectures with an intuitive algorithm for extraction face pose orientations in three dimensions: Yaw, Pitch, and Roll. Algorithm 1 represents the steps of estimating the various face poses. Algorithm 1, tasked with detecting face pose orientations, begins by initializing variables and loading essential models for facial landmark prediction. The algorithm iteratively explores video stream A, processing sub-streams with a DLIB model to detect facial regions and landmarks. Notably, it addresses concerns regarding the absence of clarity on landmark visibility and orientation restrictions.

For each frame in the sub-stream, specific facial landmarks for the eyes and nose are extracted, contributing to the calculation of Yaw, Pitch, and Roll Angles. However, concerns arise regarding the algorithm's detail, specifically the unutilized Mouth_Lms and the correlation between Yaw and Roll angles. The Pitch angle is described as 90 degrees minus the Yaw angle, a correlation at odds with real-world variations.

Algorithm 1. Detection of Face Pose Orientations

Step 1. *Input:* Facial Fiducial Points (Detected from the face images) from Video Stram A.

Step 2. *Output:* Regression Outcome (Yaw, Pitch, and Roll Positions)

Step 3. Procedure: FPOD_Mechanism ()

Step 4. Initialize Faces = [], Labels [], LandMarks = $[\mathrm{LM}_1, \mathrm{LM}_2, \ldots, \mathrm{LM}_{68}]$

Step 5. Load the face detector and facial landmark predictor

Step 6. While A is not completely explored **do**

Step 7. **For** sub-stream of length L_i **do**

Step 8. Deploy DLIB model to detect face ROI and Landmarks, also draw a bounding box

Step 9. Crop the faces

Step 10. Apply the image augmentation

Step 11. Attain tracking consequences R in frame

Step 12. **For** frames in the sub-stream **do**

Step 13. Read the frame consisting of faces and landmarks

Step 14. Extract specific facial landmarks for eyes and nose from 68 LMs

Step 15. Eye_LMs = {Left [], Right []}

Step 16. Nose_LMs = []

Step 17. Mouth_LMs = []

Step 18. $\mathrm{L\_E\_center} = \left\{\frac{x\ (\mathrm{L\_E\_LMs})}{2}, \frac{y\ (\mathrm{L\_E\_LMs})}{2}\right\}$

Step 19. $\mathrm{R\_E\_center} = \left\{\frac{x\ (\mathrm{R\_E\_LMs})}{2}, \frac{y\ (\mathrm{R\_E\_LMs})}{2}\right\}$

Step 20. $dx = \mathrm{R\_E\_Center}_{[0]} - \mathrm{L\_E\_Center}_{[0]}$

Step 21. $dy = \frac{(\mathrm{R\_E\_Center}_{[1]} + \mathrm{L\_E\_Center}_{[1]})}{2} - \mathrm{Nose}_{\mathrm{LM}_{[1]}} - \mathrm{Mouse\_LM}_{[1]}$

Step 22. $\mathrm{Yaw\_Angle}\ (\propto) = \arctan\left\{\frac{dx}{dy}\right\}$

Step 23. $\mathrm{Pitch\_Angle}\ (\beta) = \arctan\left\{\frac{dx.dx - dy.dy}{dx.dy}\right\}$

Step 24. $dx_{\mathrm{roll}} = \mathrm{L\_E\_Center}_{[0]} - \mathrm{R\_E\_Center}_{[0]}$

Step 25. $dy = dy_{\mathrm{roll}} = \frac{(\mathrm{R\_E\_Center}_{[1]} + \mathrm{L\_E\_Center}_{[1]})}{2} - \mathrm{Nose\_LM}_{[1]}$

Step 26. $\mathrm{Roll\_Angle}\ (\gamma) = \arctan\left\{\frac{dy}{dx}\right\}$

Step 27. $x_t = \mathrm{TLModels(frames\ with\ Yaw, Pitch, and\ Roll\_Angle)}$

Step 28. Add x_t to R

Step 29. **End For**

Step 30. Add R to the Final regression outcome

Step 31. **End For**

Step 32. End While

Step 33. End Procedure

As seen in Algorithm 1, we detected parameters such as center eye point and rotation angle for three different orientations that are sufficient to estimate face pose variations. These identified points help the modified transfer learning models to learn more intuitive information from these extracted transformation features. Face normalization is the process of setting each image to an appropriate range.

The deep neural network-based CNN model [18] achieves parameter reduction through the use of convolutional layers, which share weights across different spatial locations within an image. CNNs are designed to capture hierarchical features in images. CNNs are particularly effective at capturing both spatial and semantic

information in images. Convolutional layers learn spatial patterns such as edges, textures, and shapes, while higher-level layers learn semantic features such as object parts, object classes, and relationships between objects. This hierarchical representation enables CNNs to understand the content of images and make accurate predictions. Transfer learning can be applied in different ways with CNNS by leveraging the pre-trained weights (i.e., learned features and representations) that were previously trained on a specific task and later utilizing their knowledge by fine-tuning the weights for some different but related problem. Transfer learning techniques use computational resources and save time over training a complex network from scratch. Transfer learning is the most preferable method as it can boost the performance of the target task. In case of limited data problems or when training from scratch is not feasible due to some constraints, transfer learning is more reliable and convenient to handle such situations. This work utilizes the intelligence of four prominent transfer learning-based models as described in the next four sub-section.

InceptionV3

InceptionV3 is a CNN architecture developed by Google that consists of a convolutional layer, pooling layer, and fully connected layer. This version is a third updated extension of the original Inception. This version allows multiple-size kernels at different scales. This model achieves excellent performance on image classification and computer vision-related problems.

Visual Geometry Group (VGG16)

VGG16 is also a CNN-based architecture developed by the Visual Geometry Group at the University of Oxford. This model contains 16 layers, including convolutional blocks, pooling, and fully connected layers. This model has a simple and uniform architecture with only one filter size, i.e., 3×3. The performance of this model is good enough. However, VGG16 is expensive in terms of computational overhead and complex network design.

ResNet50

This network architecture is also known as residual network, developed by Microsoft Research. This model also contains a complex network architecture that comprises 50 layers along with residual connections. However, residual mapping helps in deep models. Here, skip connections are utilized that allow gradients to flow directly, reducing the vanishing gradient problem. Due to heavy layer architecture, this model allows better representation learning and generalization ability.

MobileNetV2

MobileNetV2 is a second variant of MobileNet and comes after an extension of MobileNetV1. This model represents a lightweight and simple CNN architectural design. This model is best suitable for low computational resource-based devices as it contains small model sizes and point-wise convolutions to increase model capacity.

All these models are trained on ImageNet pre-trained weights on the large datasets, which contain approx. 1000 different classes. We optimized these transfer learning models by adjusting hyperparameters based on extensive experiments performed on these four architectures to finalize the value of various parameters. The model's description, containing the features maps and other hyperparameters used in this paper, is presented in Table 1.

Table 1. Specification of the tuned parameters for proposed models

Fine-Tuned Parameters	InceptionV3	VGG16	ResNet50	MobileNetV2
Input image	299 × 299	224 × 224	224 × 224	224 × 224
CL	45 blocks (Total 48)	13 blocks (total 16)	3 in each Convolutional block	Bottleneck layers-19
FC, Dense, and Classification	05 (F, 1024, 512, 256, 3)	05 (F, 1024, 512, 256, 3)	05 (F, 1024, 512, 256, 3)	05 (F, 1024, 512, 256, 3)
Learning Rate	0.0005	0.0005	0.0005	0.0005
Kernel Size	1 × 1, 3 × 3, 5 × 5, 7 × 7	3 × 3	1 × 1, 3 × 3, 7 × 7	1 × 1, 3 × 3, 7 × 7
Batch Size	96	96	96	96
Pooling	Average	Average	Average	Average
Optimization	SGD	SGD	SGD	SGD
Dropout	0.5 (Dense)	0.25 (Dense)	0.5 (Dense)	0.25 (Dense)
Number of Epochs	300	300	300	300
Activation	ReLU	ReLU	ReLU	ReLU
Regressor	MSE	MSE	MSE	MSE
Callbacks	ReducedLR, EarlyStopping	ReducedLR, EarlyStopping	ReducedLR, EarlyStopping	ReducedLR, EarlyStopping
Total number of parameters	30,066,467	18,259,779	31,851,395	8,162,371
Trainable parameters	30,032,035	18,259,779	31,798,275	8,128,259

The process of fine-tuning involves adjusting the hyperparameters of the model after unfreezing the top layers and training the entire network with our data stream. We performed several experiments to find the best classification accuracy. Various hyper-parameters, such as learning rate, activation function, the total count of epochs, batch size, pooling, kernel size, optimizer, early stopping, and dropout were precisely fine-tuned and then finalized accordingly. The final layer of the deep network is the regression layer, where the minimum squared error loss function is utilized to estimate the different orientations of the facial poses.

4. Experiments and result analysis

The experimental setup for this work involved using an interactive interactive Python notebook (i.e., Google Colaboratory). Google Colab is an open-source cloud-based platform that provides free GPU and TPU support, regardless of the user's system configurations. During training, we utilized the “Tesla K80” GPU device accessed through CUDA version 11.2 to accelerate the preprocessing of image matrices. For programming purposes, we utilized Python 3.7.10 version along with Tensorflow 2.4.1 and Keras 2.4.3. All experiments were conducted using these tools and the resulting outcomes were evaluated and analyzed comparatively across our proposed model.

4.1. Dataset used

The AFW dataset refers to the Annotated Faces in the Wild dataset. It is a widely used benchmark dataset in computer vision and facial recognition research, especially for face pose variations problems. It is a widely used benchmark dataset that contains images that are collected from the open source (i.e., the internet) consisting of faces in different poses under unconstrained settings, making it more representative of real-world scenarios. The AFW dataset provides annotations for fiducial points of the facial attributes, such as corners of the eyes, nose, and mouth. These points are essential in the context of detecting the faces and alignment of the facial coordinates. The diversity and challenging nature involved in the AFW dataset make it a valuable resource for developing robust and accurate facial analysis systems. The AFW dataset consists of JPEG and PNG format images. However, this dataset contains varying resolution and aspect ratios. As a large data size is required for deep learning approaches, the AFW datasets include annotated images that are often stored in formats like XML, JSON, or text files. It is important to note that the configuration of the AFW dataset can be modified or customized as per the needs of the specific research problems.

Fig. 3. Samples of the AFW dataset [19]

4.2. Performance measures

To estimate the effectiveness of the proposed models, the test samples are evaluated for identifying three different orientations, i.e., yaw, pitch, and roll using a regression method. Regression is a method whose purpose is to estimate the prediction for the correlation of dependent and independent variables. Unlike the classification technique, regression techniques do not expect classification accuracy, instead, regression techniques aim at evaluating the loss function in terms of errors. By the scope of this work, we use mean absolute error to evaluate the performance of regression models.

Mean Absolute Error (MAE) [20, 21]. Mean absolute error is a measure of the average error for all variables involved in estimating the absolute difference between the ground truth value and the predicted value, which is widely used to evaluate the regression model performance. Here, a lower value indicates that the model's predictions are closer to the actual values, which implies a better fit to the data. It always shows error measures in positive terms due to absolute function,

$$\text{MAE} = \frac{1}{N}\sum|Y - Y'|, \tag{1}$$

where Y is the ground truth value Y' is the predicted value and N is the total number of data samples.

4.3. Experimental outcomes

The feature extraction and regression are performed on the AFW dataset using four dedicated deep transfer learning-based models. Table 2 represents the training losses and mean absolute errors evaluated over the AFW dataset for all four proposed models based on three face pose orientations, i.e., yaw, pitch, and roll.

Table 2. Experimental training outcome (Loss/MAE) for all the proposed models

Model	Loss	Yaw_MAE	Pitch_MAE	Roll_MAE
InceptionV3	0.0135	0.0080	0.0032	0.0030
VGG16	0.0015	1.3946×10^{-4}	2.4751×10^{-4}	0.0011
ResNet50	0.0159	0.0078	0.0040	0.0044
MobileNetV2	0.0187	0.0090	0.0054	0.0050

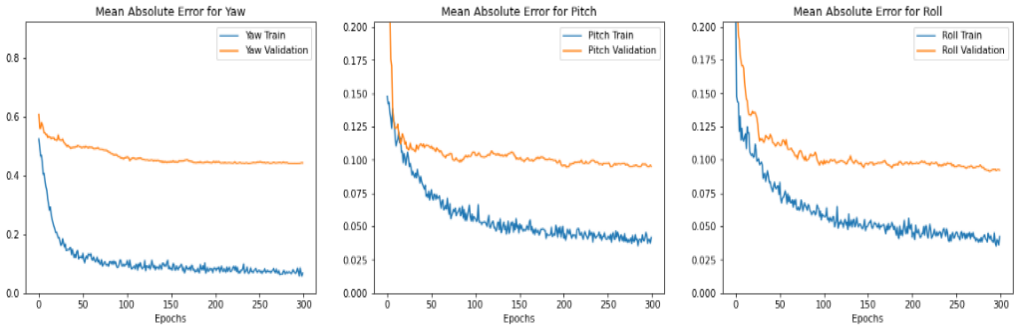
In Table 3, validation results over the AFW dataset showcase the models' effectiveness. While VGG16 maintains its excellence, InceptionV3, ResNet50, and MobileNetV2 present higher validation losses and MAE values. These findings underscore VGG16's robustness in both training and validation phases, emphasizing its potential for accurate face pose estimation across diverse applications in computer vision and biometrics.

Table 3. Experimental validation outcome (Loss/MAE) for all the proposed models

Model	Val_Loss	Val_Yaw_MAE	Val_Pitch_MAE	Val_Roll_MAE
InceptionV3	0.3786	0.3262	**0.0183**	0.0173
VGG16	**0.0937**	**0.0487**	0.0198	**0.0140**
ResNet50	0.5595	0.5138	0.0539	0.0443
MobileNetV2	0.5216	0.4084	0.0279	0.0310

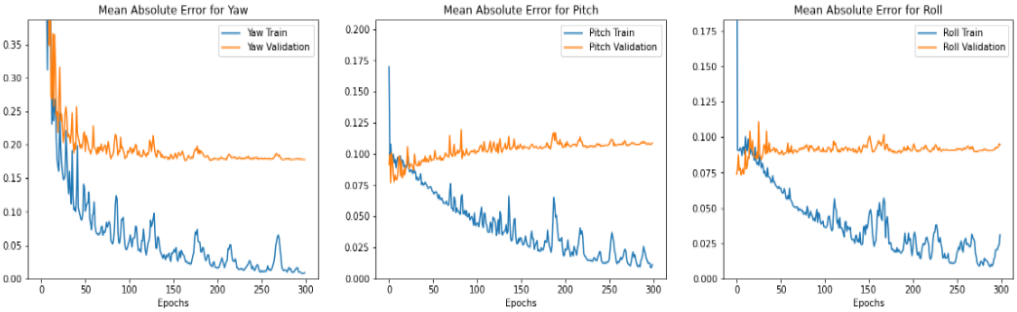
The experiment results reveal that the proposed work achieves the minimum validation loss for VGG16 whereas the same VGG16 model obtained the best validation loss for Yaw and Roll positions. The InceptionV3 performs efficiently with a minimum loss for pitch orientation. The proposed model VGG16 represents the minimum mean absolute error for Yaw and Roll orientation. In contrast, InceptionV3 shows a minimum loss in detecting the pitch orientation of the face. A trade-off between the training loss and validation loss along with mean absolute errors for all

four proposed models are presented in Fig. 4 (InceptionV3), Fig. 5 (VGG16), Fig. 6 (ResNet50), and Fig. 7 (MobileNetV2).



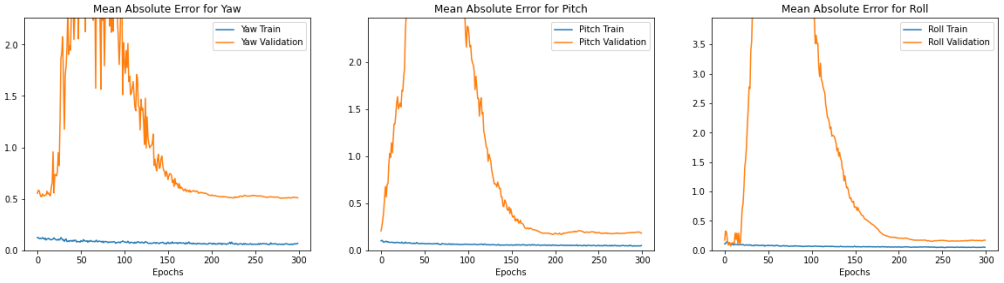
(a) (b) (c)

Fig. 4. Comparative graph between training vs validation MAE for Inception V3: Yaw (a); Pitch (b); Roll (c)



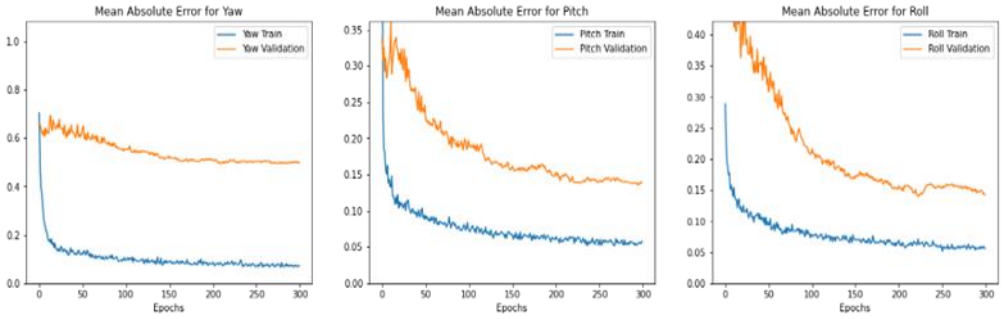
(a) (b) (c)

Fig. 5. Comparative graph between training vs validation MAE for VGG16: Yaw (a); Pitch (b); Roll (c)



(a) (b) (c)

Fig. 6. Comparative graph between training vs validation MAE for ResNet50: Yaw (a); Pitch (b); Roll (c)



(a) (b) (c)

Fig. 7. Comparative graph between training vs validation MAE for MobileNetV2: Yaw (a); Pitch (b); Roll (c)

These figures depict the performance of our proposed approach for estimating the different face pose variations in multiple dimensions, i.e., through horizontal axis, vertical axis, and longitudinal to horizontal and vertical directions. The real-time video stream capturing and subsequently, the performance analysis on the extracted feature frames is also analyzed to validate the performance of our proposed system.

4.4. Outcomes of the real-time experiments

The real-time performance analysis is performed on our proposed models. The best performance achieved by the VGG-16 model is also tested for real-time evaluations for our lab environment. We found extremely good performance in estimating the exact face pose orientation in multiple dimensions, such as yaw, pitch, and roll positions. The illustration of the real-time performance is depicted in Fig. 8.

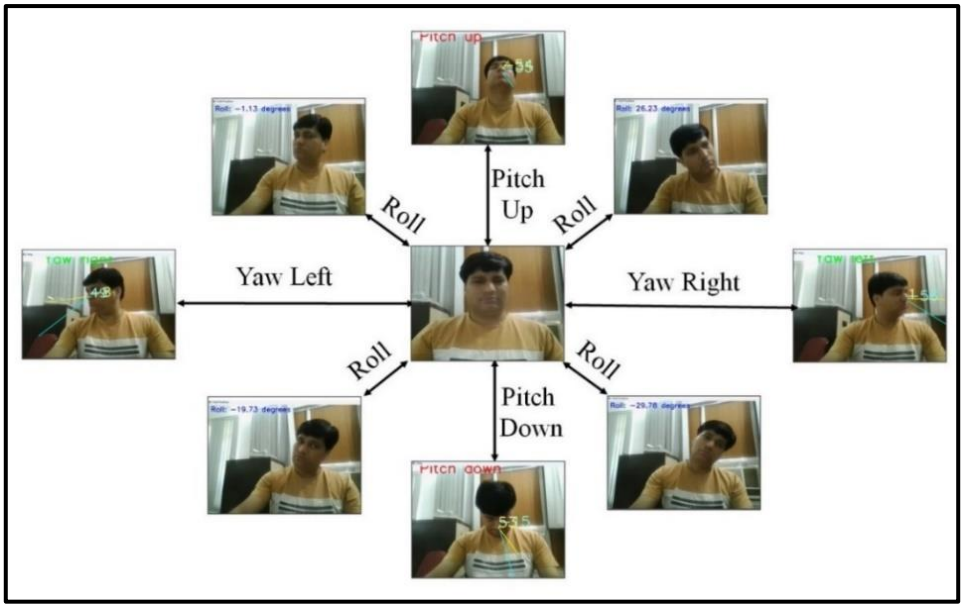


Fig. 8. Performance of our proposed approach on real-time video streams (frames) for all three orientations

4.5. Comparison with other state-of-the-art methods

The outcomes of our proposed models are compared with other state-of-the-art methods to check the effectiveness in estimating the different face poses from the proposed models. Table 4 represents the mean absolute error for different proposals underlying different architectures performed on AFW datasets compared with our two best proposals. The VGG16 presents an extremely high performance for yaw and roll positions, whereas the InceptionV3 shows the best performance for correctly estimating the pitch position. Table 4 illustrates the baseline architectures deployed in state-of-the-art works with the obtained best accuracy performed on the AFW dataset. All the considered comparisons are properly cited with their references.

Table 4. Comparison of the proposed models with state-of-the-art methods for the AFW dataset

Reference	Baseline Architecture	Mean Absolute Error (MAE)
[22]	3D Point Distribution Model and Cascaded Coupled-regressor.	9.42
[23]	Cascaded CNN regressor and 3DMM.	7.43
[24]	A special visualization layer-based CNN architecture	6.27
[25]	An Ensemble of Model Recommendation Trees (EMRT)	3.55
[26]	Efficient H-CNN Regressors (KEPLER) for key points estimation and H-CNN (Heatmap-CNN) for pose prediction.	3.01
[27]	Pose Conditioned Dendritic Convolution Neural Network (PCD-CNN)	2.36
Proposed Model	Deep Transfer Learning Model (VGG16)	0.0487 (Yaw) 0.0140 (Roll)
	Deep Transfer Learning Model (InceptionV3)	0.0183(Pitch)

5. Conclusions

This paper focuses on investigating the performance of four fine-tuned transfer learning-based models (VGG16, InceptionV3, ResNet50, and MobileNetV2) for identifying different face poses over the benchmark AFW dataset. The objective is to evaluate how facial attributes are displaced from their original (frontal face) position to new positions, which includes yaw, pitch, or roll-wise rotations or shifts. This estimation process described in the paper aims to detect faces accurately even if they are not in a frontal pose and determine their directional movement based on horizontal, vertical, and longitudinal positions. To achieve this, the paper proposes a new algorithm that combines traditional image processing operations with the detection of face regions, facial landmarks (fiducial points), and the calculation of three different orientations (yaw, pitch, and roll positions). The feature extraction process involves transfer learning consisting of InceptionV3, VGG16, ResNet50, and MobileNetV2 models with fine-tuned hyperparameters to leverage learned features of the pre-trained model. This technique improves the detection of face poses and the

estimation of facial attribute displacement in various orientations. The VGG16 reveals excellent validation performance to estimate yaw and roll positions. Whereas the InceptionV3 outperforms others to estimate pitch position. The other two models also reflect a comparative outcome. Our proposal outperforms the state-of-the-art approaches for identifying and estimating the different face poses.

Shortly, we will develop a more generalized face pose invariant model that will be capable of dealing with different unconstrained conditions, such as face poses, illumination variation, and facial expressions over more realistic datasets.

References

1. Boehm, A., D. Chen, M. Frank, L. Huang, C. Kuo, T. Lolic, D. Song. Safe: Secure Authentication with Face and Eyes. – In: Proc. of International Conference on Privacy and Security in Mobile Systems (PRISMS'2013), IEEE 2013, pp. 1-8.
2. Frischholz, R. W., A. Werner. Avoiding Replay-Attacks in a Face Recognition System Using Head-Pose Estimation. – In: Proc. of IEEE International SOI Conference. Proceedings (Cat. No 03CH37443), IEEE, 2003, pp. 234-235.
3. Rusia, M. K., D. K. Singh. A Color-Texture-Based Deep Neural Network Technique to Detect Face Spoofing Attacks. – Cybernetics and Information Technologies Vol. **22**, 2022, No 3, pp. 127-145.
4. Singh, D. K., D. S. Kushwaha. Ilut-Based Skin Color Modeling for Human Detection. – Indian J Sci Technol, Vol. **9**, 2016, No 32.
5. Singh, D. K., D. S. Kushwaha. Analysis of Face Feature-Based Human Detection Techniques. – International Journal of Control Theory and Applications, Vol. **9**, 2016, No 22, pp. 173-180.
6. Ali, Asad, et al. Gaze-Based Presentation Attack Detection for Users Wearing Tinted Glasses. – In: Proc. of 8th International Conference on Emerging Security Technologies (EST'19). IEEE, 2019.
7. Singh, M., A. S. Arora. A Novel Face Liveness Detection Algorithm with Multiple Liveness Indicators. – Wireless Personal Communications, Vol. **100**, 2018, pp. 1677-1687.
8. Rusia, M. K., D. K. Singh. A Comprehensive Survey on Techniques to Handle Face Identity Threats: Challenges and Opportunities. – Multimedia Tools and Applications, Vol. **82**, 2023, No 2, pp. 1669-1748.
9. Wu, H., K. Zhang, G. Tian. Simultaneous Face Detection and Pose Estimation Using Convolutional Neural Network Cascade. – IEEE Access, Vol. **6**, 2018, pp. 49563-49575.
10. Ranjan, R., V. M. Patel, R. Chellappa. 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, Vol. **41**, 2017, No 1, pp. 121-135.
11. Zhang, K., Z. Zhang, Z. Li, Y. Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. – IEEE Signal Processing Letters, Vol. **23**, 2016, No 10, pp. 1499-1503.
12. An, Z., W. Deng, J. Hu, Y. Zhong, Y. Zhao. APA: Adaptive Pose Alignment for Pose-Invariant Face Recognition. – IEEE Access, Vol. **7**, 2019, pp. 14653-14670.
13. He, M., J. Zhang, S. Shan, M. Kan, X. Chen. Deformable Face Net for Pose Invariant Face Recognition. – Pattern Recognition, Vol. **100**, 2020, 107113.
14. Han, Z., W. Song, X. Yang, Z. Ou. Face Pose Estimation with Ensemble Multi-Scale Representations. – In: Proc. of 2nd International Conference on Artificial Intelligence and Pattern Recognition, 2019, pp. 97-101.
15. Masi, I., F. J. Chang, J. Choi et al. Learning Pose-Aware Models for Pose-Invariant Face Recognition in the Wild. – IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. **41**, 2018, No 2, pp. 379-393.

16. F a r d, A. P., H. A b d o l l a h i, M. M a h o o r. ASMNet: A Lightweight Deep Neural Network for Face Alignment and Pose Estimation. – In: Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1521-1530.
17. Y i n, X., X. L i u. Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition. – IEEE Transactions on Image Processing, Vol. **27**, 2017 No 2, pp. 964-975.
18. L o n g, D u o n g T h a n g. Efficient DenseNet Model with Fusion of Channel and Spatial Attention for Facial Expression Recognition. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 1, pp. 171-189.
19. Z h a n g, S h i f e n g, et al. Single-Shot Scale-Aware Network for Real-Time Face Detection. – International Journal of Computer Vision, Vol. **127**, 2019, pp. 537-559.
20. R a m d a s, S., N. T. A g n e s N e e n u. Leveraging Machine Learning for Fraudulent Social Media Profile Detection. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 1, pp. 118-136.
21. A l-D u j a i l i, M. J., A. H. J a b a r S a b a t. A New Hybrid Model to Predict Human Age Estimation from Face Images Based on Supervised Machine Learning Algorithms. – Cybernetics and Information Technologies, Vol. **23**, 2023, No 2, pp. 20-33.
22. J o u r a b l o o, A., X. L i u. Pose-Invariant 3D Face Alignment. – In: Proc. of IEEE International Conference on Computer Vision, 2015, pp. 3694-3702.
23. J o u r a b l o o, A., X. L i u. Large-Pose Face Alignment via CNN-Based Dense 3D Model Fitting. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 4188-4196.
24. J o u r a b l o o, A., M. Y e, X. L i u, L. R e n. Pose-Invariant Face Alignment with a Single CNN. – In: Proc. of IEEE International Conference on Computer Vision, IEEE, 2017, pp. 3200-3209.
25. Z h u, S., C. L i, C. C. L o y, X. T a n g. Towards Arbitrary-View Face Alignment by Recommendation Trees, 2015, arXiv preprint arXiv:1511.06627.
26. K u m a r, A., A. A l a v i, R. C h e l l a p p a. Kepler: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient h-CNN Regressors. – In: Proc. of 12th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2017, pp. 258-265.
27. K u m a r, A., R. C h e l l a p p a. Disentangling 3D Pose in a Dendritic CNN for Unconstrained 2D Face Alignment. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 430-439.

Received: 30.05.2023; Second Version: 04.01.2024; Third Version: 07.04.2023;
Accepted: 21.04.2024