

# Human and Machine Vision Based Indian Race Classification Using Modified-Convolutional Neural Network

Vani A. Hiremani\* and Kishore Kumar Senapati

Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, 835215, Jharkhand, India

\*Corresponding Author: Vani A. Hiremani. Email: vani.hiremani@gmail.com

Received: 21 January 2022; Accepted: 06 April 2022

**Abstract:** The inter-class face classification problem is more reasonable than the intra-class classification problem. To address this issue, we have carried out empirical research on classifying Indian people to their geographical regions. This work aimed to construct a computational classification model for classifying Indian regional face images acquired from south and east regions of India, referring to human vision. We have created an Automated Human Intelligence System (AHIS) to evaluate human visual capabilities. Analysis of AHIS response showed that face shape is a discriminative feature among the other facial features. We have developed a modified convolutional neural network to characterize the human vision response to improve face classification accuracy. The proposed model achieved mean F1 and Matthew Correlation Coefficient (MCC) of 0.92 and 0.84, respectively, on the validation set, outperforming the traditional Convolutional Neural Network (CNN). The CNN-Contoured Face (CNN-FC) model is developed to train contoured face images to investigate the influence of face shape. Finally, to cross-validate the accuracy of these models, the traditional CNN model is trained on the same dataset. With an accuracy of 92.98%, the Modified-CNN (M-CNN) model has demonstrated that the proposed method could facilitate the tangible impact in intra-classification problems. A novel Indian regional face dataset is created for supporting this supervised classification work, and it will be available to the research community.

**Keywords:** Data collection and preparation; human vision analysis; machine vision; canny edge approximation method; color local binary patterns; convolutional neural network

## 1 Introduction

Inter-class classification problems like classifying Indian face vs. Chinese face [1] are pretty feasible than intra-class classification problems like Indian face vs. Indian face [2]. This problem is more apprehensive in a highly populated and diversified country like India, where every region epitomizes different cultures and traditions. Over the years, humans have showcased clever proficiency in judging age, gender, behavior, state of mind, and race by face, even under many obstacles [3–6]. The human brain processes visual statistics in semantic space by extracting the semantically imperative features such



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

as contour information, line segments, edges which are hard to detect by computers. In contrast, machines process them in data space obtained by the strongly detectable but less informative features like texture patterns and chromatic information [7]. Knowing how a human extracts the features can empower diversified artificial intelligence applications employing human-like performance. To explore this, we have proposed an AHIS involving untrained identifiers selected at random in a fine-grained race classification problem to evaluate the potential of human vision. The interrogation of identifiers on given regional face images showed the influence of local conventional facial features like face shape, skin tone, eyebrows, the shape of eyes, orientation of mouth, shape of the nose, and non-conventional features like the style of dressing, vermilion color, and style, draping sari as per regional tradition, physic, mustache, accessories like jewelry and regional amulet thread. This rich feature set will be a reference input set to train neural models for solving computer vision problems. The insight of non-conventional features definitely would suffice the absence of conventional features. Augmentation of these symmetric features can bring tangible gain in classification. The experimental results of human visual analysis have shown an accuracy of 88% when both identifiers and persons in the image are from different regions [8]. 96% accuracy is achieved when both are from the same region [9]. This familiarity of faces [10] reinforced improved performance in classification. The proficiency of humans in underlying classification problems is systematically measured, and the derived discriminative features are characterized through computer vision algorithms using a novel face database. This work emphasizes the strength of CNNs since they have achieved commendable success in image classification on large-scale datasets for a long time [11]. CNN's are made up of neurons that have learnable weights and biases. They compare the image patch by patch, typically  $3 \times 3$  or  $5 \times 5$  matrices. Each neuron receives some inputs performs a convolution operation between a  $3 \times 3$  patch of image and  $3 \times 3$  filters. The strength of CNN is that they learn local patterns of images rather than global patterns like densely connected layers do. ConvNets learn spatial hierarchies of patterns progressively from abstract to complex features layer by layer. Since patterns are translation invariant, the neuron in a layer is only connected to a small portion of the layer ahead of it. This mechanism reduces the number of computing parameters. Upfront the success of the AlexNet has tremendous influence in the classification process employing sinking filter size [12] or escalating the network deepness [13]. CNN model training is a global optimization problem. We proposed different variations to improve the traditional CNN model to achieve the best fitting parameters by three aspects: Inception module, spectral pooling, and leaky ReLu activation function. The comparative analysis of proposed M-CNN against conventional computational models has shown approximately 92.98% accuracy. Alongside to explore the perceptual annotation of individual feature influence in the overall face, face contour information is obtained from the canny edge detection approximation method and characterized through the CNN model. To this end, a large set of regional face database is created to address the scarcity of region-wise labeled Indian face databases. This database is made public for further research work in addition to relatively available few datasets [14,15] of Indian faces with no regional information.

The main contribution of the paper is as follows:

- We have developed an automated human intelligent system incorporating 120 identifiers to estimate human vision proficiency. Both conventional and non-conventional features are recorded, which will be prospected as a reference input set to train neural models for solving computer vision problems.
- We propose variations to the conventional CNN model based on the empirical experiments by including the Inception module, spectral pooling, and leaky ReLu activation function. The statistical hypothesis testing and comparative analysis have shown that the proposed M-CNN model outperformed other conventional models in accuracy and execution time.

- We present the substantial perceptual annotation of individual features influence in the overall face by obtaining face contour information through canny edge detection approximation method and characterized through CNN model.
- We have developed a novel Indian regional face database consisting of 2895 faces acquired from north, east, west, and south via online and offline mode. It is labeled database emphasizes a supervised classification problem. It will be available for the research community.

The rest of the paper is organized in the following sections: Section 2 details the proficiency of both humans and machines in race classification in the past. A broad elaboration of materials and methods used in this research work is in Section 3. Section 4 explains the evaluation metrics and analysis of results attained from both humans and machines. Section 5 concludes our work to an extent and details the scope of future work.

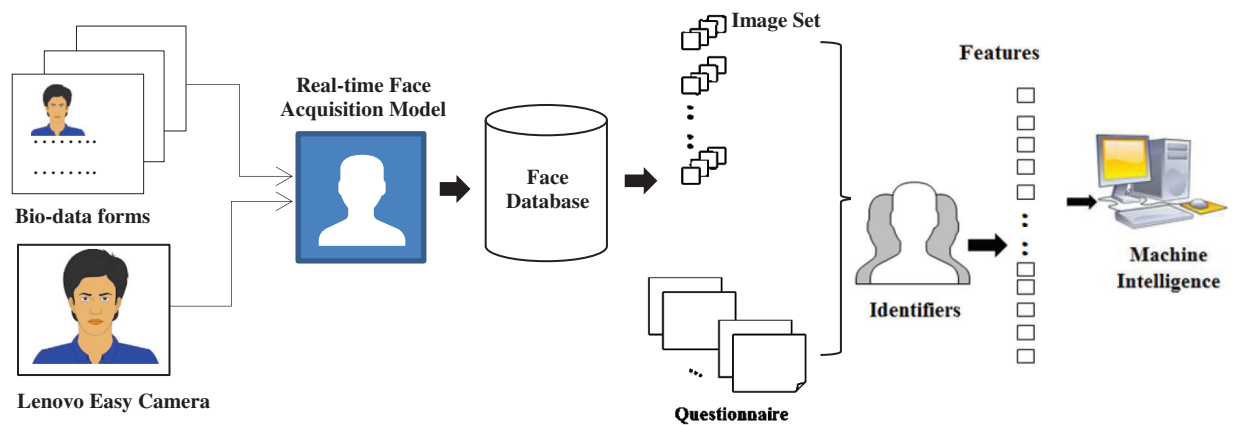
## 2 Related Works

Geographical regional faces do have a stereotyped structure with many discriminative features. Humans and machines systematically process them, applying learning experience and computational logic. In 2009, the racial classification of Caucasian, Black, and Asian abrasive races and gender [16,17] was performed precisely using computer vision. In [18], silhouetted face profiles are given to human identifiers exhibiting a lot of ethnicity information, and based on shape and color, gender is decided. Classification problem becomes handy with subtle feature variations, i.e., finer grained race such as Chinese/Japanese/Korean [19], Chinese sub-ethnicities [20], and Myanmar [21]. In this section, the different inter and intra-class classification problems are discussed, and how human vision has always been influential on machine intelligence in solving computer vision problems is seen. Classification of coarse races such as Caucasian/Black is performed at both human and machine sides with 70%–80% accuracy. Humans can reliably judge the region of a person based on skin complexion, the way the person behaves, facial makeup, accessories, or fashion sense like dressing style and hairstyle. Like humans visualize the input image, divide it into distinct regions, extract the area of interest out of the whole image and validate positioned face, the machine also pursues the image in the same way. It visualizes low-resolution images using low pass filters, segments them into the distinct region using a segmentation algorithm, and extracts a region of interest (ROI) from the image using an object detecting algorithm. Finally, the given face image is verified [22]. In [23], Caroline E. Harriott et al. have proven that more or less human and machine practices the same tactics to do a job by pairing human-machine and human-human participants to find a suspicious thing in an artificial setup. In [24], Kun Yu et al. have developed a human-computer interaction system to help physically challenged people. The way human learns face gestures based on facial features a set of interactive gestures are designed using Face++ to train a machine to achieve interaction with the computer. In [25], Zhihao Shen et al. have trained the machine to boost human-robot communication by inferring the human traits such as eyesight, body gesture, energy, pitch, and Mel-Frequency Cepstral Coefficient (MFCC) through human-robot interaction. In current intelligent manufacturing systems, the human-machine interaction process has become the most crucial aspect, extending to autonomous systems where human trust dynamics can be utilized to improve the human-machine interaction process. Computational models like Local binary patterns, Gabor filters bank, and wavelets are feature extraction schemes performed with excellence like humans [26]. Features play a very vital role in computer vision problems. Selecting the most informative features yields radical improvements in the classification rates. Artificial intelligence is explored to mimic the tasks of human brains in solving science and engineering problems [27]. CNN is being used on a large scale to train machines with massive samples for feature learning. CNN is at the top and outperformed other computational models in classifying Chinese, Korean and Japanese faces.

The survey has described the influence of human intelligence in solving computer vision problems like face recognition, gender classification, object detection, sentiment analysis, and many more. However, these problems fall under inter-class problems where umpteen numbers of features exist to classify the input. The state-of-the-art unveils scarcity of intra-class classification where limited discriminant features are available. This work has encouraged us to address this challenge by proposing a model incorporating humans and machines in racial classification.

### 3 Materials and Methods

The human and machine-centric face classification architecture is depicted in Fig. 1. It consists of 3 principal parts: (1) Database creation, (2) Automated human intelligence system, and (3) Machine intelligence system (MIS).



**Figure 1:** The proposed model describes three parts: 1. Face images are acquired through static and dynamic mode to create Indian face database. Pre-processing is handled at this stage (only cropping is performed under pre-processing). 2. Each identifier is interrogated against a set of 10 face images along with questionnaire form. 3. Computational models are used to characterise the feature and the approach used by human to classify human face to particular region

#### 3.1 Data Pre-processing and Annotation

A novel Indian color face database is created to mitigate the scarcity of regional and labeled face images for the underlying supervised classification process. We have sought permission from many universities running in the east and south regions. After receiving the consent, the face images of faculties, staff, students, and their family members are collected through two modes: online (real-time user interface model) and offline (Bio-Data forms with consent disclaimer). The face images are acquired from various states belonging to the east and south regions. An automated face acquisition model is developed to capture real-time face images under candidates' consent. To capture images, we have used a Lenovo Easy Camera of 2 mp with an aspect ratio of 1.33 and a resolution of size  $640 \times 480$ . The candidate is positioned in front of the Lenovo easy camera, the undecorated room wall being in the background. The region of interest (ROI) is detected through the Viola jones algorithm and captured to the size  $250 \times 350$ . In offline mode, the scanned Bio-data forms containing candidate photos with minimal information are collected from various universities. These segmented images are then browsed through the automated face acquisition model where the unnecessary background and labeled information are removed, and only ROI is captured to the size of  $250 \times 350$ . Around 2010, images acquired from these modes are stored with the primary key as Region\_Number (e.g., EAST\_01). Images are pre-processed beforehand instead

of direct feeding to CNN. A one-hot encoded vector is generated from the categorical name of images. The dependent variables, i.e., labels, are encoded for machine understanding as the dataset consists of categorical names (i.e., SOUTH\_01 and EAST\_01). The dataset consists of varying size images, so the different resolution images are reduced to the size 50 \* 50 pixels and converted into grayscale [28] images to curtail processing speed. 80% of images are considered for training, and the remaining 20% for testing. The following Tab. 1 shows the summary of the dataset split. The Train\_Set and Test\_Set images are reshaped to size (-1, 50, 50, 1) to fit in TensorFlow.

**Table 1:** Summary of splitting of dataset into training and testing sets

| Dataset         | South | East | Total |
|-----------------|-------|------|-------|
| Train_Set (80%) | 1158  | 1158 | 2316  |
| Test_Set (20%)  | 290   | 289  | 579   |
| Total           | 1448  | 1447 | 2895  |

### 3.2 AHIS Model: Classification Task

This section describes an empirical examination carried on humans to analyze and understand which features they considered to classify given faces to their regions. Let S be a human intelligence system consisting of input image  $I_i$  (picked from  $I_{2895}$  (face database)), questionnaire Q (set of 9 questions, i.e.,  $Q = \{q_1, \dots, q_9\}$ ), feature vector Fe ( $Fe = \{f_1, f_2, \dots, f_L\}$ ) [8] extracted by human identifiers, answer A (the subset of features Fe in terms of answers,  $A \subseteq Fe$ ), and C the result of binary classification ( $w_1$  and  $w_2$ ). The representation of AHIS(S) is described as  $S = \{I_i \mid I_i \in I_{2895}, Q, Fe, A, C\}$ . For the given classification task of C and unknown patterns represented by feature vector  $F_e$ , we computed conditional probability P as  $P(w_i \mid Fe)$ , where i represents two classes. After all instant represents the probability that the unknown pattern belongs to the respective class  $w_i$ , given that the feature vector incorporates the features from Fe. Let's say  $w_1$  and  $w_2$  are the two classes consisting of expected patterns. The priori probability  $P(w_1)$  and  $P(w_2)$  are estimated from the available training feature vectors. Suppose N is the total available training pattern and instance  $N_1, N_2 \subseteq N$ . If  $(N_1, N_2)$  belong to  $(w_1, w_2)$  respectively then,  $P(w_1) \approx N_1/N$  &  $P(w_2) \approx N_2/N$ . The classification now can be stated as,

If  $P(w_1 \mid Fe) > P(w_2 \mid Fe)$ , Fe is classified to  $w_1$

If  $P(w_1 \mid Fe) < P(w_2 \mid Fe)$ , Fe is classified to  $w_2$

Let  $R_1$  be the region of the feature space which we decide in favor of  $w_1$  and  $R_2$  be the corresponding region for  $w_2$ . Then error is made if  $F \in R_1$ , although it belongs to  $w_2$ , or if  $F \in R_2$ , although it belongs to  $w_1$ . That is,

$$P_e = P(F \in R_2, w_1) + P(F \in R_1, w_2)$$

$P_e$  is the joint probability of the  $P(F \in R_2, w_1)$  and  $P(F \in R_1, w_2)$  events. Fe's outstanding features are utilized for training the model for further classification. In AHIS, a sample of 120 (shown in Tab. 2) untrained identifiers is selected randomly from various regions of India. Each candidate knows the motto of their contribution beforehand the interrogation to upholding ethical and participatory research.

#### 3.2.1 Human Interrogation

During this phase, the randomly considered identifiers are questioned to their intellect. As shown in Fig. 1, around 2316 images among 2895 are arranged in 122 sets. Each identifier is given a set consisting of  $\pm 19$  images. Each identifier is interrogated with a set of fundamental questions [8] and images. The

questions are framed to record how the identifier perceives an image, the gender, the discriminating features, and any additional factors those favored identifiers to guess the region.

**Table 2:** Demography of identifiers with respect to age parameter

| Groups | Quantity of group | Age   |
|--------|-------------------|-------|
| 1      | 60                | 20–30 |
| 2      | 60                | 30–40 |

### 3.2.2 Feature Analysis

The digitally signed filled forms are meticulously evaluated in this phase. Every identifier has found images adequate for identification, despite challenging images. Despite the absence of any regional accessories, identifiers still decided the correct region looking upon facial features like light complexion, space between eyebrows with small eyes, lightly shaded eyebrows, and marginally sunken cheekbones in case of extreme east regions face. Each identifier's 10 questionnaire forms are evaluated thoroughly and recorded in an excel sheet incorporating identifiers region, information of images given to them, identified facial features, area of the image identified, its validation, and overall accuracy. The records collectively revealed many facts to human proficiency in identifying the Indian regional people. Two significant observations are made: 1. Identifiers have observed not only conventional face features but also considered non-conventional features. 2. The factor of belonging to the same region played a significant role [9]. 3. The identifiers who mentioned face shape as a promising feature to classify were unable to express what they meant by face shape. Few answers suggested that it is not the shape but the face's aura or depiction of look that classifies them to a particular region.

To address this limitation at human side the machine vision is explored to characterise this face shape.

### 3.3 MIS Model: Classification Task

This section describes the computational models like Color local binary pattern (CLBP) and CNN. We improved the feature extraction scheme, Local Binary Patterns (LBP), by adding color factor to it. CLBP is used to comprehend the features observed by identifiers. According to [27], neural networks are best suited for mimicking human intelligence. Therefore we have built two CNN models, CNN-FC and M-CNN. The CNN-FC model is trained with 2316 contoured images obtained from the canny edge approximation method. The M-CNN model also trained with 2316 standard face images to characterize the overall features drawn at the human side.

#### 3.3.1 Color Local Texture Features Extraction

Feature extraction is a crucial step in every computer vision problem. Due to discriminative power and computational simplicity, the LBP texture operator is considered a more stature feature for face recognition. The critical application of the LBP operator is its robustness to monotonic grayscale changes caused (i.e., illumination variations). This LBP operator is applied to the color face image to transform it into a CLBP image. The CLBP image is blocked into 256 cells ( $16 \times 16$ ), i.e., each cell consists of  $8 \times 8$  pixels resolutions. Initially, the space structure of the face is reserved. Then for each square, the LBP histogram is calculated to statistically reflect the edge sharpness, flatness of region, existence of unique points, and variety of local region attributes. Then to each block of the color image LBP function is applied [28]. The LBP feature vector is the concatenated serial of all  $16 \times 16$  intensity values computed from the histogram generated using Y, R, G, and B color components of individual instances of the images. Therefore, the LBP feature is a statistical texture description of the image consisting of a series of

histograms of blocked sub-images. The feature dimensions are determined by blocking number and the sampling density. Hence, the size of the template for 100 users is  $10000 \times 256$ . During the construction of the LBP feature vector, the bilinear interpolation method is adopted on the LBP grid to estimate the values of neighbors that do not fall precisely on pixels. Since the correlation between pixels decreases with distance, more texture information is obtained from local neighborhoods. We have considered a 300 face image dataset consisting of 5 different images of 60 people from the east and south region. 4 images out of five are trained, and one is tested. The global features are also preserved to retain ample scale information of the image to avoid deformation of images of a person. The CLBP features of trained images are matched against CLBP features of a testing image using Manhattan distance-based algorithm.

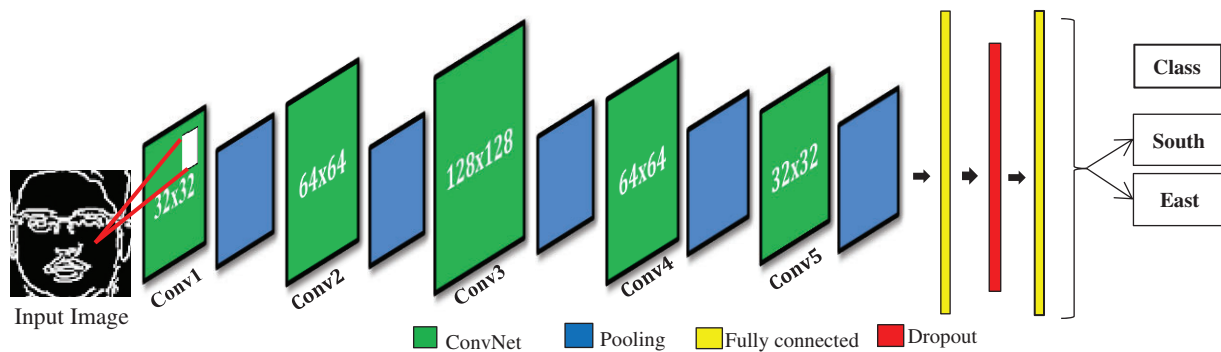
### 3.3.2 CNN-FC

CNN's are the type of feed-forward Artificial Neural Networks. The output of such a neural network for any input pattern  $z_p$  is calculated with a single forward pass-through network in Eq. (1). For each output  $O_k$ , we have (assuming a few hidden layers in between the input layer and output layer),

$$\begin{aligned}
 O_{k, p} &= f_{O_k}(\text{net}_{O_k, p}) \\
 &= f_{O_k} \left( \sum_{j=1}^{J+1} \cdot w_{kj} f_{y_j}(\text{net}_{y_j, p}) \right) \\
 &= f_{O_k} \left( \sum_{j=1}^{J+1} \cdot w_{kj} f_{y_j} \left( \sum_{i=1}^{I+1} \cdot v_{ji} z_{i, p} \right) \right)
 \end{aligned} \tag{1}$$

where  $f_{O_k}$  and  $f_{y_j}$  are the active functions for output  $O_k$  and hidden layer  $y_j$ ,  $w_{kj}$  is the weight between output  $O_k$  and  $v_{ji}$ ,  $z_{i, p}$  is the value of input  $z_i$ , the  $(I+1)^{\text{th}}$  input unit and the  $(J+1)^{\text{th}}$  hidden unit are bias units representing the threshold values of neurons in further layer to adjust the weight.

The essential components of CNN-FC are (1) Convolutional Layers, (2) ReLU Layers, (3) Pooling Layers, (4) Fully connected layers, and (5) Dropout layer. Fig. 2 represents the architecture of CNN-FC.



**Figure 2:** Constellation of CNN-FC architecture used to classify countered face images to region

#### Convolution Layers

In Fig. 2, the CNN-FC model has five convolutional layers. The first layer consists of  $32 \times 32$  filters, the second layer of  $64 \times 64$  filters, the third layer of  $128 \times 128$  filters, the fourth layer of filter size  $64 \times 64$ , and the fifth layer of  $32 \times 32$  filters. The traditional CNN has a fixed size of filters at different levels, and usually, filter size tends to decrease. Here we have used different filters of sizes  $32 \times 32$ ,  $64 \times 64$ , then  $128 \times 128$  with stride five, because filters are generally related to feature maps that will be flattened at the end to distinguish

more shapes and textures. It is good practice to use filters of multiples of 2. It provides better RAM organization and batch size. At  $(m, n)$  location in the  $x^{\text{th}}$  feature map of  $y^{\text{th}}$  layer,  $z_{i, j, k}^1$ , the feature value is calculated by Eq. (2).

$$Z_{m,n,y}^y = w_x^{yT} X_{m,n}^y + b_x^y \quad (2)$$

where  $w_x^y$  and  $b_x^y$  are the weight vector and bias term of the  $x^{\text{th}}$  filter of the  $y^{\text{th}}$  layer, respectively and  $X_{m,n}^y$  is the input patch centered at location  $(m, n)$  of the  $y^{\text{th}}$  layer.

### ***ReLU (Rectified Linear Unit) Layers***

CNN adapts to learn non-linear data. Most of the real-world data samples learned are primarily non-linear. Since the convolution layer is linear in operation, the ReLU layer helps convert the linear process to non-linear. ReLU transformation function  $f(x)$  is used to activate the nodes if the input  $(x)$  is above the threshold value, while if the input is below zero, then the output is zero. It showed a linear relationship with the dependent variable as in Eq. (3).

$$f(x) = \begin{cases} 0 & \text{if } X < 0 \\ X & \text{if } X \geq 0 \end{cases} \quad (3)$$

It has removed every  $-ve$  values from the filtered image and turns them to 0's. Its non-saturation of gradient quality makes it a good choice in CNNs.

### ***Pooling Layers***

In this layer, we performed shift-variance by reducing the images obtained from the ReLU layer into a smaller size. After every layer, the feature map is halved without compromising the information. The feature map of each pooling layer is associated with the preceding layers corresponding feature map. Eq. (4) computes the pooling function  $\text{pool}(\cdot)$  for each feature map.

$$y_{m,n,k}^1 = \text{pool}\left(a_{m,n,k}^1\right), \forall(m, n) \in R_{i,j} \quad (4)$$

where  $R_{i, j}$  represents a local neighborhood around  $(i, j)$  location. We have used 5 max pooling layers with  $5 \times 5$  windows.

### ***Fully Connected Layers***

These are the final layers where the high-level reasoning takes place. The filtered and shrunken images are put in a single list. Two fully connected layers are used: one is of 1024 neurons, and the other is of 2 neurons.

### ***Dropout Layer***

This regularization technique is used for avoiding overfitting by preventing co-adaptations on Train\_Set. A single dropout layer is added with a key probability of 0.8 ( $p = 0.8$ ), followed by a dual-node decision layer. The output of this layer is denoted as in Eq. (5).

$$O = bv * a (W^Z i) \quad (5)$$

where  $i = [i_1, i_2, \dots, i_n]^Z$  is considered input to fully-connected layer,  $W \in M^{p \times q}$  denotes weight matrix, and  $bv$  represents a binary vector of size  $q$  being a Bernoulli distribution with parameter  $p$ , i.e.,  $r_i \sim \text{Bernoulli}(p)$  as a source of every element. Finally, to reduce cross-entropy loss, the Adam optimizer is used with a learning rate  $\alpha = 0.001$ .

This model is now trained with contoured face images obtained by the given algorithm.



**Algorithm: Canny Edge Detection Approximation**

Input: Initially, the contour information of ROI is extracted from face image with respect to centroid using  $C_x = M_{10}/M_{00}$  and  $C_y = M_{01}/M_{00}$ .

Step 1: Apply the Gaussian filter shown in Eq. (6) to remove the noise.

$$S = G\sigma * I \quad G\sigma = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6}$$

Step 2: Compute the intensity gradient of image (x, y) using Eq. (7).

$$\nabla S = \left[ \frac{\partial}{\partial x} S \quad \frac{\partial}{\partial y} S \right]^T = [S_x \quad S_y]^T \tag{7}$$

Step 3: Apply non- maximum suppression technique to avoid the fake edges (i.e., the pixels in gradient magnitude images are suppressed that are not of local maximum).

Step 4: Apply double threshold to handle remaining spurious responses. (i.e., edge pixels with weak gradient value)

Step 5: Finally edges are tracked by Hysteresis (i.e., considering only promising edges with high value and suppressing weaker ones)

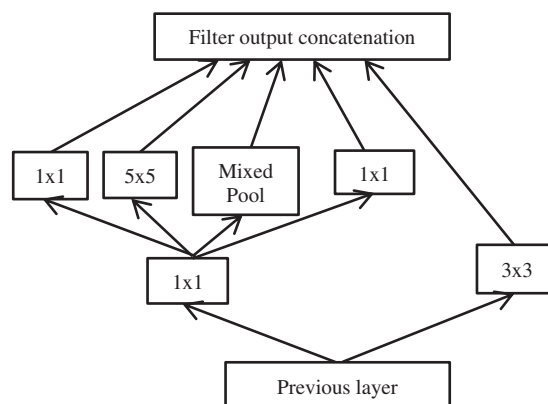
The contour approximation method ensured that all the image points were stored, keeping the original image intact.

**3.3.3 M-CNN**

Based on the empirical experiments, we have proposed variations to the CNN model by incorporating: The inception module, spectral pooling, and leaky ReLu activation function.

**Inception Module**

In conventional CNNs, convolution filter is a generalized linear model (GLM) representing the input image area. It is more suitable for the samples where abstract features are linearly separable. Here we propose an improved inception module to enhance its representation ability. These are used in CNN's to reduce computational complexity and decrease the deeper network's dimensionality with stacked  $1 \times 1$  convolutions. Instead of having either  $3 \times 3$  or  $5 \times 5$  filters or a pooling layer, this model suggests having all of them. As shown in Fig. 3, the new architecture incorporates all  $1 \times 1$ ,  $3 \times 3$ ,  $5 \times 5$  filters, and mixed pooling layers. The convolutional operation is performed on every output of the previous layer. The concatenated output from all filters passed as an input to the next layer. This process allowed the increase in depth and width of CNN without increasing the computational complexity.



**Figure 3:** Improved inception model

In the first step, we have applied 128 filters of different sizes  $1 \times 1$  and  $3 \times 3$  on an input image. The feature map obtained from this step is fed to the second step so that all filters of different sizes  $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$  and mixed pooling should perform on the same image. Padding is kept identical to maintain the same output and input shape of Conv2D operation. So the outcome of each filter is the same. It helps in concatenating the output of each filter to get the output of the inception module. Such modules can solve the computational expense and overfitting issues.

### Mixed Pooling

The function in Eq. (8) represents the mixed pooling technique. Here we have combined both max pooling and average pooling to have a better solution for the overfitting problem instead of applying alone of them.

$$Y_{i,j,k} = \lambda_{(m,n) \in Ri,j} \max a_{m,n,k} + (1 - \lambda) \frac{1}{|Ri,j|} \sum a_{m,n,k} \quad (8)$$

where  $\lambda = \{0, 1\}$  indicates the choice of max pooling or average pooling, respectively, the two-dimensional window runs over each channel of an input image, and a filter covers the features lying within the region. A feature map (FM) of dimension  $M \times N \times K$  gets a new size shown in Eq. (9) after the pooling layer.

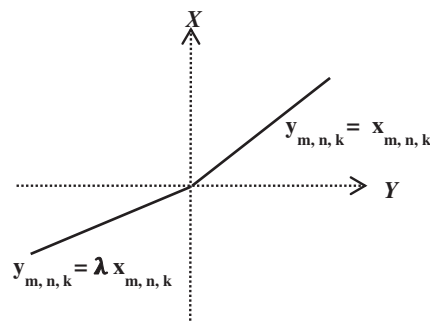
$$FM = \frac{(M - f + 1)}{s} \times \frac{(N - f + 1)}{s} \times K \quad (9)$$

where  $M$  and  $N$  are the height and width of the feature map, respectively,  $K$  is the number of channels,  $f$  is the size of the filter, and  $s$  is stride length.

### Activation Function

A potential disadvantage of ReLU function  $X_{m,n,k} = \max(y_{m,n,k}, 0)$  found in CNN-FC is that the gradient turned to be zero whenever the unit was inactive. This disadvantage causes the problem of gradient-based optimization for weight adjustment. The training process is downturned because some idle units have never been active due to the persistent zero gradients. We incorporated leaky ReLU (Fig. 4) defined in Eq. (10) to alleviate this problem.

$$X_{m,n,k} = \max(y_{m,n,k}, 0) + \lambda \min(y_{m,n,k}, 0) \quad (10)$$



**Figure 4:** Leaky ReLU function,  $\lambda$  is default parameter value ranges between 0 and 1

Unlike ReLU, the negative part in Leaky ReLU is compressed rather than mapping it to constant zero, which results in yielding a small and non-zero gradient while the unit is idle.

### ***Loss Function Regularization and Optimizer***

The underlying binary classification (where the number of classes  $C = 2$ ) found that the cross-entropy loss shown in Eq. (11) is more suitable since it measures the classification model performance whose likelihood output value falls between 0 and 1.

$$\text{Loss} = \sum_{i=1}^2 y_i \log p_i = - [y \log (p) + (1 - y) \log (1 - p)] \quad (11)$$

where  $y_i$  is positive information ranging within 0 and 1 and  $p_i$  is  $i^{\text{th}}$  class Softmax probability. We have flattened the output to a 1-D array of neurons fed to two fully connected layers one is of 1024 neurons, and the other is of 2 neurons corresponding to two classes (decision layer). Overfitting occurred by preventing co-adaptations on Train\_Set is reduced using the dropout regularization technique. A single dropout layer with 0.8 ( $p = 0.8$ ) key probability is added, followed by a dual-node decision layer. Finally, the model is compiled with Adam optimizer to update weights iterative based in Train\_Set with learning rate  $\alpha = 0.001$ . The feature vector for each face consists of 1024 features.

## **4 Evaluation Metrics and Result Discussions**

### **Development Environment**

We have used Microsoft Windows 10 operating system as a primary system requirement with 2 GHz CPU processing speed and 4 GB of RAM. We installed Anaconda Navigator, an open-source distribution of python to implement computational models. The CNN and M-CNN models are developed in Spyder (the scientific python development environment. Additionally, we used TensorFlow with a GPU notebook provided by Google colab on a Linux-based hosted machine.

### **4.1 Analysis of AHIS and MIS**

We estimated the proficiency of identifiers rigorously on two Indian databases and compared the performance of computational models on the proposed database. The similar correlations are obtained for the two datasets ( $r = 0.68 \pm 0.05$  for Set 1,  $r = 0.61 \pm 0.02$  for Set 2; correlation in Set 1 > Set 2 in 885 of 1000 random samples). Set 1 and Set 2 have achieved 58% and 59.4% accuracy, respectively, based on assumptions made on regions. The main limitations with available relative databases are that they do not consist of region-wise labeled faces and are not adequate for supervised learning. The proposed database addresses this issue. The performances, including precision, recall, and F1 score of the proposed M-CNN model, are evaluated based on True Positive (TP), False Positive (FP), and False Negative (FN) metrics for the novel labeled dataset. All the metrics calculated are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

According to the Mann-Whitney U test, the probability of samples correctly classified from the South population is more remarkable than samples from the East population and is different (larger or smaller) than the probability of samples from the East exceeding the samples from the South; i.e.,  $P(\text{South} > \text{East}) \neq P(\text{East} > \text{South})$  or  $P(\text{South} > \text{East}) + 0.5 \cdot P(\text{South} = \text{East}) \neq 0.5$ . Since the south region faces have more accurately classified (with correlation  $r = 79.5\%$ ) than the East (approximately  $r = 74.3\%$ ). The M-CNN model has shown a high stratified rate for shabby images (overall 77.9%) compared to a single featured based classifier and traditional CNN (61% and 65%, respectively). Tab. 3 presents a rich feature set drawn from the analysis of human visual response. It comprises both conventional and non-conventional facial features.

**Table 3:** Conventional and Non-conventional features observed by identifiers

| Region | Local conventional features  | Non-conventional feature  |
|--------|--|---|
| North  | Skin tone: fair, light brown<br>Nose: straight, sleek<br>Eyes: normal size<br>Face shape: oval, sleek  | Loud make up, hair style, salwar suit dress, prominent application of vermilion, sari draping style, appearance and authentic jewellery.  |
| East   | Skin tone: light brown, dark brown<br>Nose: wider, prominent in face, tiny nose, flatten nose bone, Broad forehead<br>Eyes : puffy, hooded eyelids, tiny eyes, small eyebrows, significant distance between eyebrows<br>Lips: fuller, small<br>Face shape: round, bulgy, double chin square, smooth skin texture, prominent cheek bone | Saffron and pink coloured vermilion applied from head till nose found occasionally, head covered with sari, sari draping style, amulet thread colour tied on arms. Red colour thread found around married women neck. Typical and very minimal jewellery. |
| West   | Skin tone : fair light brown, Nose: small, Eyes : normal size  | Jewellery, vermilion, sari style, studs   |
| South  | Skin tone: fair, light brown, dark brown<br>Eyes: normal size, dark eyebrows, Face shape: round, oval  | Traditional vermilion and jewellery, flower, hair style, sari style, moustache  |

The accuracy of the proposed model is measured using the Genuine Acceptance Rate (GAR) and False Acceptance Rate (FAR) [28] performance evaluation metrics as stated in Eqs. (12) and (13) below:

$$\text{GAR} = \frac{\text{Correctly classified faces}}{\text{Total face images}} \times 100\% \quad (12)$$

$$\text{FAR} = \frac{\text{Wrongly classified faces}}{\text{Total face images}} \times 100\% \quad (13)$$

#### 4.2 Comparative Analysis of M-CNN against Conventional CNN using Statistical Hypothesis Test

This section performed a comparative analysis based on the proposed Indian regional face dataset. The performance of the proposed M-CNN model against the conventional CNN model is measured using the Chi-Square statistical hypothesis testing. Let's say M-CNN represents M1 and CNN is M2. Tabs. 4 and 5 presents the confusion matrixes of both M1 and M2.

**Table 4:** M1 observation table

| Model M1(M-CNN) | Predicted(s) | Predicted(e) |
|-----------------|--------------|--------------|
| Actual(s)       | 267          | 23           |
| Actual(e)       | 23           | 266          |

\* s = south region face, e = East region face, Number of South region faces actual(s) vs. predicted(s) = 290, Number of East region faces actual(s) vs. predicted(s) = 289. The predicted parameter is calculated based upon the accuracy (i.e.,  $267/290 \approx 92.06\%$  and  $290-267 = 23$ ).

**Table 5:** M2 observation table

| Model M2(CNN) | Predicted(s) | Predicted(e) |
|---------------|--------------|--------------|
| Actual(s)     | 247          | 43           |
| Actual(e)     | 43           | 246          |

Number of South region faces actual(s) vs. predicted(s) = 290, Number of East region faces actual(s) vs. predicted(s) = 289. The predicted parameter is calculated based upon the accuracy (i.e.,  $247/290 \approx 85.2\%$  and  $290 - 247 = 43$ ).

**Chi-square test for evaluating M-CNN**

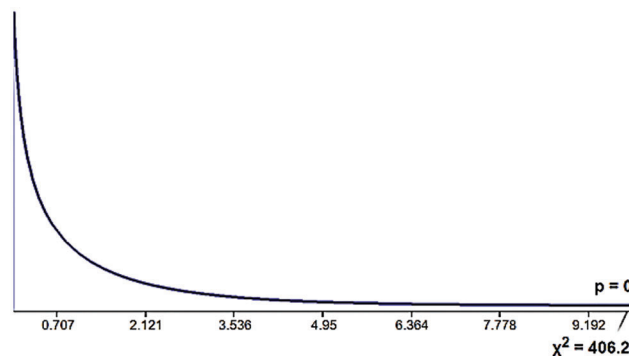
According to the M1 observation table, the probability of data instances belonging to South is  $\pm 92.06\%$  and a  $\pm 7.9\%$  chance of East otherwise. In chi-square tests, we extract the expected values from observations. M1 (M-CNN) labels 290 instances as South. If M1 is randomly guessing, we can expect approx. 7.9% of those instances to be of East. Since there is 7.9%, the chances are that a test data instance is East. Hence according to the law of independent probability given in Eq. (14),

$$P(\text{Predicted} = s \text{ and Actual} = e) = P(\text{Predicted} = p) \times P(\text{Actual} = a) \tag{14}$$

We can derive the value of  $P(\text{Predicted} = s \text{ and Actual} = e)$ .  $P(\text{Predicted} = s) = 290/100 = 2.9$

$$P(\text{Predicted} = s \text{ and Actual} = e) = 2.9 \times 0.079 = 0.23$$

So, 23% of the total data instances are likely to be classified as East face. Therefore, the number of East faces = 23% of 100 = 23. The following Fig. 5 shows the chi-square distribution with degree of freedom (DOF = 1).



**Figure 5:** Chi-square distribution (DOF = 1) for Chi-square statistic = 406.26 P value = 2.4e-90

The chi-square distribution graph shows that the chi-square statistic is exceedingly high, and the probability (p-value) of a null hypothesis is insignificant compared to the alpha (0.05). Thus, we can claim that M1 is not a random predictor and better fits the data.

**Comparing M-CNN against CNN using Matthew Correlation Coefficient (MCC)**

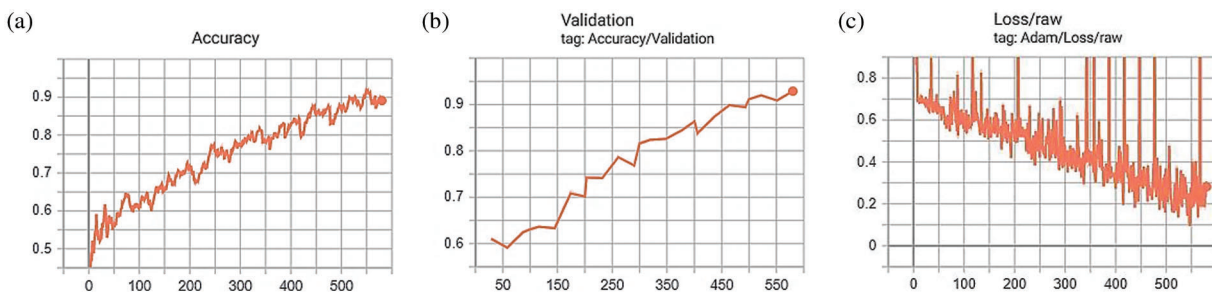
We have used MCC statistics to evaluate the M1 and M2 models performance. Based upon the confusion matrices of M1 and M2 shown above, we have calculated the critical classification metrics such as 1. Accuracy (how suitable M1 is at prediction). 2. Sensitivity (how often M1 chooses the positive class when the observation is in the positive class). 3. Precision (how often an M1 is correct when it predicts

the positive class). 4. Specificity (how often M1 chooses the negative class when the observation is a negative class). [Tab. 6](#) presents the performance of M1 and M2 based on the metrics mentioned above.

**Table 6:** M1 (M-CNN) performance report

| Measure                          | Derivations  | Value         | Value         |
|----------------------------------|--|---------------|---------------|
| Sensitivity                      | $TPR = TP/(TP + FN)$   | <b>0.9207</b> | <b>0.8517</b> |
| Specificity                      | $SPC = TN/(FP + TN)$   | <b>0.9204</b> | <b>0.8512</b> |
| Precision                        | $PPV = TP/(TP + FP)$   | <b>0.9207</b> | <b>0.8517</b> |
| Negative Predictive Value        | $NPV = TN/(TN + FN)$   | 0.9204        | 0.8512        |
| False Positive Rate              | $FPR = FP/(FP + TN)$   | 0.0796        | 0.1488        |
| False Discovery Rate             | $FDR = FP/(FP + TP)$   | 0.0793        | 0.1483        |
| False Negative Rate              | $FNR = FN/(FN + TP)$   | 0.0793        | 0.1483        |
| Accuracy                         | $ACC = (TP + TN)/(P + N)$  | <b>0.9206</b> | <b>0.8515</b> |
| F1 Score                         | $F1 = 2TP/(2TP + FP + FN)$   | <b>0.9207</b> | <b>0.8517</b> |
| Matthews Correlation Coefficient | $TP * TN - FP * FN / \sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}$ | <b>0.8411</b> | <b>0.7029</b> |

The recorded values against different metrics show that M1 significantly outperforms M2, the conventional CNN used in phase 3.3.2 in terms of accuracy, sensitivity, specificity, F1-score, and Matthew correlation coefficient. The Tensorboard graphs shown in [Fig. 6](#) depict the accuracy and loss of model M1 (M-CNN). [Fig. 6a](#) chart shows the accuracy of 89.07% against the testing set. [Fig. 6b](#) shows the validation of the model as 92.98% against the training set. [Fig. 6c](#) demonstrates the cross-entropy loss estimation.



**Figure 6:** Visualization of feature mapping of training set images to testing image set using Tensorboard graphs. (a) Accuracy of model drawn from training the training set images with respect to image batch, (b) Validation of model with respect to testing image set and, (c) Loss by cross entropy

[Tab. 7](#) presents the cumulative performance of human vision and the different computational models discussed in this work.

**Table 7:** The records the accuracy of different models used in proposed work

| Model  | Train_Set | Test_Set | FAR  | GAR   |
|--------|-----------|----------|--|---|
| AHIS   | –         | 2316     | 4% in case of same region<br>14% in case of different region | 96% in case of same region<br>88% in case of different region |
| CLBP   | 840       | 260      | 23%  | 67%   |
| CNN    | 2316      | 579      | 14%  | 85.07%  |
| M-CNN  | 2316      | 579      | 10%  | 92.98%  |
| CNN-FC | 2316      | 579      | 31%  | 70.7%   |

## 5 Conclusion

This research investigated human and machine intelligence performance under the Indian race classification problem involving human identifiers and computational models like CNN, M-CNN, CNN-FC, and CLBP. This work has shown the significance of customizing Deep Learning models for specific applications and projects. For intra-class face classification problems, the proposed M-CNN is highly recommended compared to the traditional CNN model. The CNN-FC model trained with contoured face images has characterized the influence of face shape among other facial features. This model would be prospected to explain the impact of face shape information. We have developed a novel Indian regional face database to mitigate the scarcity of labeled databases for underlying supervised classification problems. The experimental outcome showed the high proficiency of humans (96%) compared to machine vision algorithms (92.98%).

**Acknowledgement:** The authors acknowledge BIT Mesra, Ranchi for providing necessary tools and journal access in need. We are grateful to the renowned universities for providing help in data collection.

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- [1] Y. Wang, Y. Feng, H. Liao, J. Luo and X. Xu, "Do they all look the same? Deciphering Chinese, Japanese and Koreans by fine-grained deep learning," in *Proc. IEEE Conf. on Multimedia Information Processing and Retrieval*, Miami, FL, USA, pp. 39–44, 2018.
- [2] H. Katti and S. Arun, "Are you from North or South India? A hard race classification task reveals systematic representational differences between humans and machines," *Journal of Vision*, vol. 19, no. 1, pp. 1–17, 2018.
- [3] P. Sinha, B. Balas, Y. Ostrovsky and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.
- [4] P. Sinha, B. Balas, Y. Ostrovsky and R. Russell, "Face recognition by humans: 20 results all computer vision Researchers should know about," Department of Brain and Cognitive Sciences Massachusetts Institute of Technology, Cambridge, MA, pp. 1–26, 2005.
- [5] R. Chellappa, P. Sinha and P. Jonathon Phillips, "Face recognition by computers and humans," *Computer*, vol. 43, no. 2, pp. 46–55, 2010.
- [6] K. Jain, B. Klare and U. Park, "Face recognition: Some challenges in forensics," in *Proc. IEEE Int. Conf. on Automatic Face & Gesture Recognition*, Santa Barbara, CA, USA, pp. 726–733, 2017.
- [7] B. Zhang, "Computer vision vs. human vision," in *Proc. IEEE Int. Conf. on Cognitive Informatics*, Beijing, China, pp. 3–9, 2010.

- [8] V. Hiremani and K. Senapati, "Analysis of human intelligence in identifying persons native through the features of facial image," in *Proc. IEEE Int. Conf. on Advances in Information Technology*, Chikmagalur, India, pp. 87–93, 2019.
- [9] V. Hiremani and K. Senapati, "Significance of conventional and nonconventional features in classification of face through human intelligence," in *Proc. IEEE Int. Conf. on Smart Systems and Inventive Technology*, Tirunelveli, India, pp. 131–137, 2019.
- [10] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy *et al.*, "Indian movie face database: A benchmark for face recognition under wide variations," in *National Conf. on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, Jodhpur, India, pp. 1–5, 2013.
- [11] K. Nogueira, O. A. Penatti and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, no. 2, pp. 539–556, 2017.
- [12] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. European Conf. on Computer Vision*, Zurich, Switzerland, pp. 818–833, 2014.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. on Learning Representations*, San Diego, CA, USA, 2015.
- [14] R. Sharma and M. S. Patterh, "Indian face age database: A database for face recognition with age variation," *International Journal of Computer Applications*, vol. 126, no. 5, pp. 21–27, 2015.
- [15] G. Somanath, M. Rohith and C. Kambhamettu, "VADANA: A dense dataset for facial image analysis," in *Proc. IEEE Int. Conf. on Computer Vision Workshops*, Barcelona, Spain, pp. 2175–2182, 2011.
- [16] K. R. Brooks and O. S. Gwinn, "No role for lightness in the perception of black and white? Simultaneous contrast affects perceived skin tone, but not perceived race," *Perception*, vol. 39, no. 8, pp. 1142–1145, 2010.
- [17] S. Fu, H. He and Z. Hou, "Learning race from face: A survey," *IEEE Transactions on PAMI*, vol. 36, no. 12, pp. 2483–2509, 2014.
- [18] U. Tariq, Y. Hu and T. S. Huang, "Gender and ethnicity identification from silhouetted face profiles," in *16th Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Cairo, Egypt, pp. 2441–2444, 2009.
- [19] X. -D. Duan, C. -R. Wang, X. -D. Liu, Z. -J. Li, J. Wu *et al.*, "Ethnic features extraction and recognition of human faces," in *Proc. IEEE Int. Conf. on Advanced Computer Control*, Shenyang, China, pp. 125–130, 2010.
- [20] H. K. Tin and M. M. Sein, "Race identification for face images," *Proc. ACEEE International Journal on Information Technology*, vol. 1, pp. 35–37, 2011.
- [21] V. Bruce, "Research article, influences of familiarity on the processing of faces," *Perception*, vol. 15, no. 4, pp. 387–397, 1986.
- [22] H. Mo, W. Li and L. Dai, "Research of face location system based on human vision simulations," in *Proc. IEEE Intelligent Computation Technology and Automation*, Changsha, China, pp. 170–174, 2008.
- [23] C. E. Harriott, G. L. Buford, T. Zhang and J. A. Adams, "Human-human vs. human-robot teamed investigation," in *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction*, Boston, MA, USA, pp. 405, 2012.
- [24] K. Yu and J. Yin, "The study of facial gesture based on facial features," in *Proc. IEEE Int. Conf. on Intelligence and Safety for Robotics*, Shenyang, China, pp. 105–109, 2018.
- [25] Z. Shen, A. Elibol and N. Y. Chong, "Inferring human personality traits in human-robot social interaction," in *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction*, Daegu, Korea (South), pp. 578–579, 2019.
- [26] J. O'Toole, P. Jonathon Phillips, F. Jiang, J. Ayyad, N. Penard *et al.*, "Face recognition algorithms surpass humans matching faces over changes in illumination," *Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1642–1646, 2007.
- [27] F. Li and Y. Du, "From AlphaGo to power system AI: What engineers can learn from solving the most complex board game," *IEEE Power and Energy Magazine*, vol. 16, no. 2, pp. 76–84, 2018.
- [28] S. A. Angadi, S. M. Hatture and V. Hiremani, "Real time face recognition for user authentication using color local texture features," in *Proc. Int. Conf. on Emerging Research in Computing, Information, Communication and Applications*, Bangalore, India, pp. 715–721, 2013.