ORIGINAL RESEARCH

Age Detection in a Surveillance Video Using Deep Learning Technique

S. Vasavi[1] · P. Vineela[1] · S. Venkat Raman[2]

Received: 10 August 2020 / Accepted: 25 March 2021

Abstract

The video streams that are collected from CCTV surveillance camera can be used in many applications such as crowd analysis, forensic, self-profile analysis, and social network user's analysis. Soft biometrics such as age, gender, height, skin color can be used for human analysis. This requires object detection and feature analysis. Works reported for face recognition and age detection have poor performance with real-world profile images. It may be because of incomplete description of the human object. Also, they rely on traditional image processing algorithms that extract hand-crafted features. Deep learning workflow transforms the identified patterns into mathematical modeling that can be used for subsequent prediction. Residual networks can skip connections and can address vanishing gradient problem with improved accuracy. Wide ResNet 34-based system is proposed in this paper that automatically predicts age of human object in video images. Modified Wide ResNet is used for feature extraction that learns facial keypoints, image reconstruction using Simultaneous algebraic reconstruction technique for up sampling, 101 number of classes (101-way classification) ranging from 0 to 100. Proposed system accuracy is evaluated using mean absolute error and with Pearson correlation coefficient that finds correlation between actual age and predicted age. Experimental results proved that data augmented Wide ResNet out performs the existing age prediction methods with 5% increase in accuracy.

Keywords Age prediction · Image reconstruction · Wide ResNet · Feature extraction · Data augmentation

Introduction

Soft biometrics can be used to know about an individual such as age, gender, hair color, etc. The following Fig. 1 presents a summary of soft biometrics and its applications. Retina scanner can be used to get the human retinal output. Finger print scanner can be used to know the edges of the fingers that is widely used in many security applications. Deep learning algorithms use neural network architectures that extract features more accurately than machine learning algorithms. This paper uses wide ResNet network for age prediction. Separate classification module attached to the network is used to predict the age. Appa Real dataset is used to evaluate the proposed model. The proposed model is evaluated for accuracy, precision recall, correlation, distribution graph, mean absolute error.

Motivation

Automatic identification of video surveillance is primarily taking an interest in undertaking a challenging project in artificial intelligence and machine learning. The problem of predicting age from an image has potential uses of this application include uses in social media, age restriction to access videos or at self-checkout grocery stores. Figure 2 presents the application of detecting the age. As the lady passes through the surveillance camera, age and gender are detected, and subsequently, the relevant advertisements are played in the way she traveled.

This article is part of the topical collection "Data Science and Communication" guest edited by Kamesh Namudri, Naveen Chilamkurti, Sushma S. J. and S. Padmashree.

✉ S. Vasavi
vasavi.movva@gmail.com

P. Vineela
potlurivineela101@gmail.com

S. Venkat Raman
sarma6019@gmail.com

[1] Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada, India

[2] Department of Space, Advanced Data Research Institute, Hyderabad, India

Published online: 30 April 2021

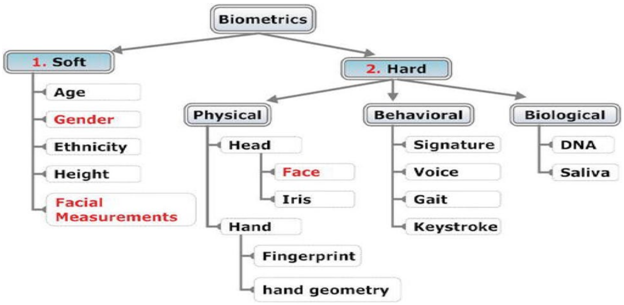


Fig. 1 Soft biometrics [29]

Contribution

The main objective of this study is to predict the age of a human object. A modified Wide ResNet is proposed in this paper and is trained on APPA Real dataset and our college faculty dataset to investigate on:

(i) A regression model that identifies the relationship between real and predicted age.
(ii) The importance of adding invariant feature plane in BN and max pool in the classification layer. Rotation-invariant features help in improving detection accuracy of object with different appearances.
(iii) Image reconstruction using Simultaneous algebraic reconstruction technique (SART) to overcome the problem of sampling.

This combined system could detect and predict small objects present in the video image with its usefulness in several business applications dependent on surveillance video of human objects.

Organization

This paper is organized as follows: Sect. 2 describes literature survey on object detection and classification methods. Section 3 describes the proposed framework. Results and discussion is given in Sect. 4.

Literature Survey

Many works are reported for age detection such as Quividi [1] for monitoring the demographics of users passing by in railway station or in super markets and to engage them with relevant brand items. Performance evaluation is done on the algorithms used in Quividi by USA company McGladrey [2] and proved with 93% accuracy. Android app such as Age-Bot can detect age of multiple objects present in an image. Convolutional Neural Network (CNN)-based models such as Caffe, Tensor flow and Torch are developed for detecting age of a human being. ResNet model to predict gender age and race is reported in [4]. Authors used Image Net dataset. They could not provide enough discriminative information. Inertial sensor-based methodology is explained in [5]. OU-ISIR GAIT dataset is used to test their approach. Their work showed performance of 86.6% for gender and age prediction. Mobile phone-based data are used to classify soft biometrics of 46 subjects in [6]. Data collection is done using PUSH and PULL components. The age and gender features are extracted. Gender, age, marital status whether they are male or female or not. The dataset used is Sherlock dataset. Intra-person variation refers to the variation that occurs within a given individual over different periods of time. Early human growth is characterized by wide intra-individual variability in body size. Body weight prediction using body contour is reported in [7]. REDDIT posts are used to test their system. Tal Hassner and Gil Levi [8] developed models to predict age and gender of a single face. CNN-based age and gender prediction is reported in [9]. IMDb-WIKI and OIU-Adience dataset are used for experimentation. Cascade classifiers and Caffee pre-trained models are used to predict the age and gender of human object recorded by webcam [18]. PhotoAgeClock system is discussed in [19] using Xception DNN-based model. 8 classes for age group prediction using VGG16 model is discussed in [20]. Authors of [17] studied on eight-connected and four-connected shift residual networks. These neighborhood shifts are applied on ImageNet for reducing the bottleneck in original ResNet as shown in Fig. 3a, b.

Fig. 2 Use case for age detection [3]

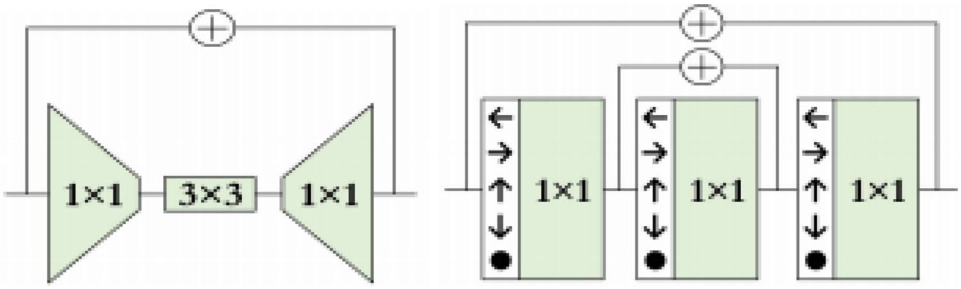


Fig. 3 **a** Original residual block design. **b** Simplified channel-flattened multi-shift residual block [17]

Very less literature is available on shift operations in wide residual networks. Many systems are reported on APPA Real Dataset such as [24–28] that applied face detection techniques followed by Deep learning architectures for age prediction. Existing works have the following drawbacks:

(i) They ca not give sufficient information.
(ii) Limited accuracy
(iii) Intra-person is characterized by individual body size
(iv) Cannot correlate between body heft and optic body
(v) Global and local features have not been distinguished for age estimation.
(vi) Training the classifiers with hand-crafted features that may sometimes leads to more time complexity and overfitting problems
(vii) Treating age prediction as Binary classification rather than multiclass classification

The proposed system uses deep learning framework that consists of wide residual neural networks with convolutional layers. ResNet blocks can be used to grow width of the network and decreases network depth, called wide residual networks (WRNs).

Proposed System

The BN-ReLU-conv reported in [14] is modified by adding an invariant feature plane, face score threshold in the BN layer, max pool in the convolution classification layer. During the process of down sampling, we may understand on "what" and may lose "where" of an image. Segmentation network of U-Net that uses transposed convolution layer for up sampling is shown in Fig. 4a, and an improved proposed system architecture with data augmentation layer and image reconstruction using SART [21] as shown in Fig. 4b is used in the proposed system. Kernels/filters are used in CNN for edge detection in an image. The proposed system detects key facial landmarks using dlib detector [16] that uses ensemble of regression trees. Mix-up augmentation that uses weighted linear interpolation for invariant feature plane. Non-linear activation function such as ReLU (Rectified Linear Unit) is used for learning kernel weights [10].

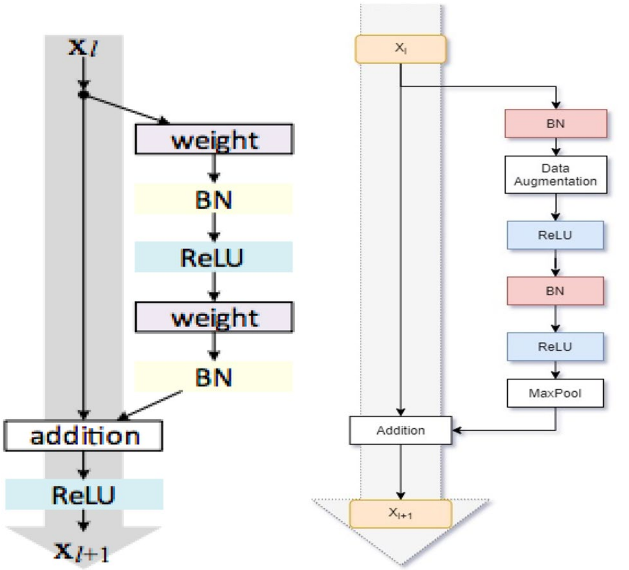


Fig. 4 **a** Original BN-ReLU-conv [20]. **b** Proposed system

Figure 5 presents the process flow diagram.

Methodology

1. Input the image
2. Pre-process the image

   (a) RGB to gray scale conversion
   (b) Resize the image 224*224

3. Train the modified WIDE_RESNET[12] model

   (a) Set the arguments (weight file, depth, width, margin, target, image, directory)
   (b) Import the weight file
   (c) Load the weights
   (d) Calculate first and second face score using isinf (facescore)

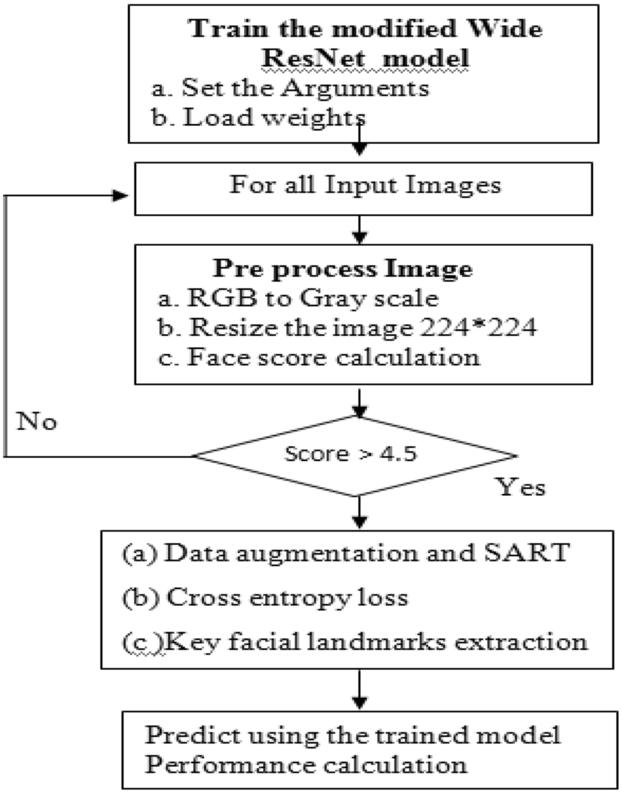

Train the modified Wide ResNet model
a. Set the Arguments
b. Load weights
For all Input Images
Pre process Image
a. RGB to Gray scale
b. Resize the image 224*224
c. Face score calculation
No
Score > 4.5
Yes
(a) Data augmentation and SART
(b) Cross entropy loss
(c)Key facial landmarks extraction
Predict using the trained model
Performance calculation

Fig. 5 Process flow diagram

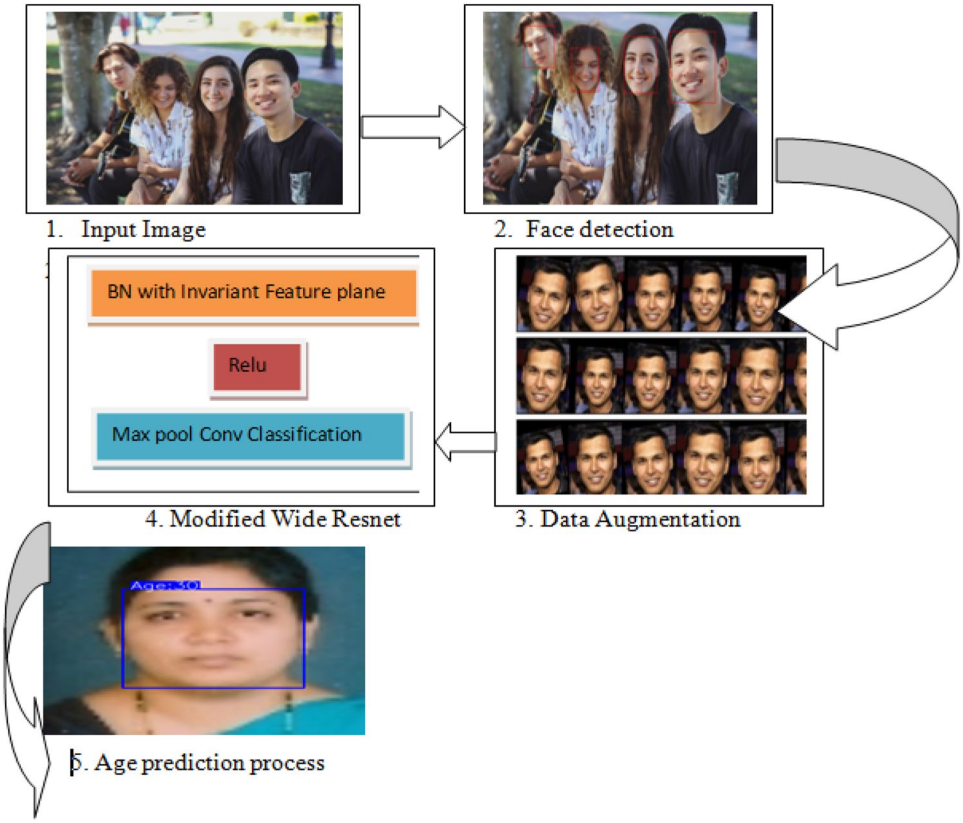
(e) For the images face score > 4.5:
 (i) Data augmentation using mix-up.
 (ii) Sampling using SART
 (iii) Calculate cross-entropy loss
 (iv) Pass each of the ROI blobs through the network for feature extraction and obtain key facial landmarks using dlib detector

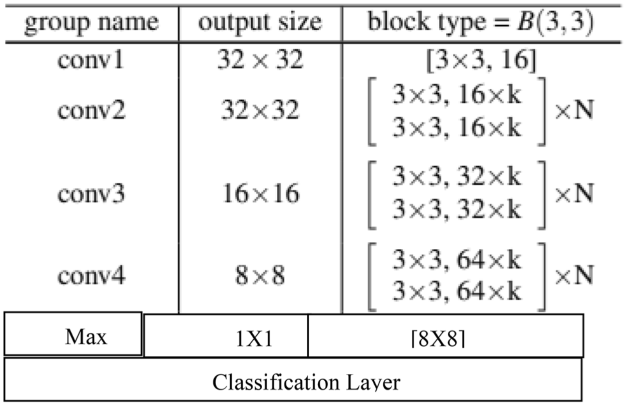
4. Predict the age using the wide ResNet model for a particular face.
5. Performance calculation: estimate accuracy.

Architecture

Even though deep residual networks scale to several thousand layers for improving performance, they fail in feature reuse and have slow training times. Also, some blocks may not help in feature learning [11]. Hence, this proposed system used varied wide residual network that has few layers, fast training time and with improved accuracy. Data augmentation is done using mix-up model [14] to improve validation loss, uses linearly interpolating between data points. Since age prediction is considered as a regression problem, Relu is used to solve vanishing gradient problem, Maxpool in the final classification layer to improve performance. Figure 6 presents the age prediction process. Figure 7 presents batch size in each of the layer in the wide ResNet model. Max pool

Fig. 6 Age prediction process





group name	output size	block type = $B(3,3)$
conv1	32×32	$[3\times3, 16]$
conv2	32×32	$\begin{bmatrix} 3\times3, 16\times k \\ 3\times3, 16\times k \end{bmatrix} \times N$
conv3	16×16	$\begin{bmatrix} 3\times3, 32\times k \\ 3\times3, 32\times k \end{bmatrix} \times N$
conv4	8×8	$\begin{bmatrix} 3\times3, 64\times k \\ 3\times3, 64\times k \end{bmatrix} \times N$
Max	1X1	[8X8]
Classification Layer		

Fig. 7 WideResNet model with Max pool layer

is added on top of classification layer. Residual networks structure used in the present study consists of Convolutional layer conv1 that is followed by three groups (each of size *N*) of residual blocks conv2, conv3 and conv4, followed by max pooling and final classification layer. Sampling using SART with 100 iterations is made, so as to identify object of size (height/width) smaller than 16 pixels. IoU is set to 0.3.

Algorithms

Algorithm for Pre-process the Image

(a) RGB to gray scale conversion.
  Grayscale = (R + G + B)/3.

(b) Resize the image as given in step (b) to the size of 224*224.
  Input tensor = Input (shape = (224, 224, 3)).

Algorithm for Training the Modified Wide ResNet Model

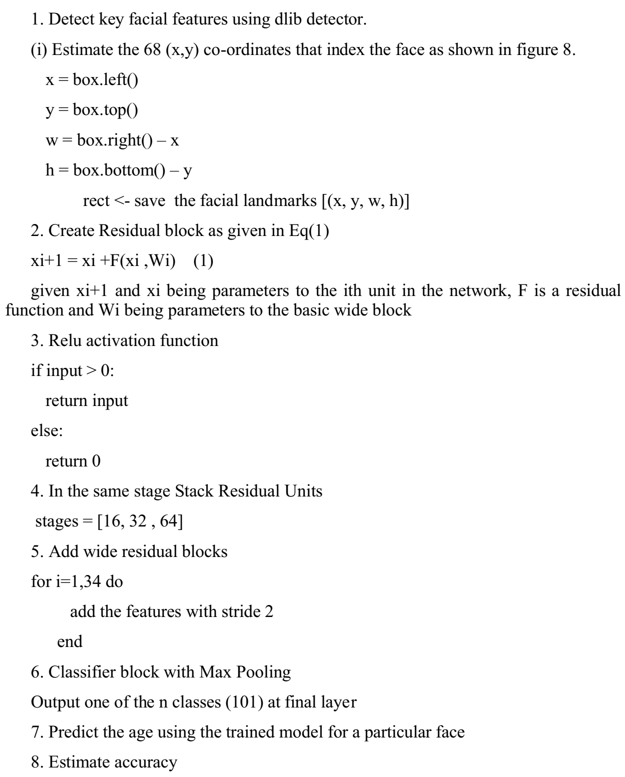
```
1. Set the parameters
num_classes = 100
num_filters = 64
num_blocks = 4
num_sub_blocks = 2
use_max_pool = False
 (b) Repeat for each face in the image
    (i) Face score calculation and elimination basing on threshold
 first_face_score = format(np.isinf(face_score).sum()))
                        second_face_score = format((~np.isnan(second_face_score)).sum()))
                          for k in range(len(first_face_score) || len(second_face_score))
     (ii) Save the images only if score >4.5,
if first_face_score [k] < 4.5 || second_face_score [k] < 4.5
                           img_paths.append(full_path[i][0])
                           show_imgs(img_paths)
(iii) For each of the saved image as in step (ii) do data augmentation
```

Algorithm for Data Augmentation Using Mix-Up for Invariant Features

For each saved image from 3.4.2 algorithm perform data augmentation as described in [15]:

```
For each saved image from 3.4.2 algorithm perform data augmentation as described in
[15]:
      1. Load image 1 and image 2 and their targets from the dataset
      2. Create new image with linear interpolation, for t value between 0 and 1 derived
from beta distribution, Beta(t; t).
       newx = t * image1x + (1-t) * image2x
       newy = t * image1y + (1-t) * image2y
      3. Sampling using SART
      self.angles = [0,15,30,45,60,75,90,105,120,135,150,165,180]
      self.iterations = 100
      4. Calculate loss L using cross entropy of these new examples, for randomly sampled
t value between 0 and 1 and pick the example with less loss.
      L = log_loss(y, y_pred)
      5. If L is acceptable (nearest to zero), Append the new example to the dataset
```

Algorithm for Feature Extraction and Prediction (Fig. 8)



```
1. Detect key facial features using dlib detector.
(i) Estimate the 68 (x,y) co-ordinates that index the face as shown in figure 8.
  x = box.left()
  y = box.top()
  w = box.right() – x
  h = box.bottom() – y
       rect <- save  the facial landmarks [(x, y, w, h)]
2. Create Residual block as given in Eq(1)
$x_{i+1} = x_i + F(x_i, W_i)$    (1)
given xi+1 and xi being parameters to the ith unit in the network, F is a residual function and Wi being parameters to the basic wide block
3. Relu activation function
if input > 0:
  return input
else:
  return 0
4. In the same stage Stack Residual Units
 stages = [16, 32 , 64]
5. Add wide residual blocks
for i=1,34 do
     add the features with stride 2
   end
6. Classifier block with Max Pooling
Output one of the n classes (101) at final layer
7. Predict the age using the trained model for a particular face
8. Estimate accuracy
```

Dataset Collection

Two datasets are used for experimentation. APPA REAL database [13] contains 7591 images, from which 121 images with multiple unclear faces removed from the dataset, giving 7470 images. Random function is used to split data these images into 4059 train set, 1472 validation set and 1939 test set. Another dataset is created with 40 college faculty. Figures 9 and 10 presents sample input and predicted images from both datasets [13].

Data augmentation is done to overcome over fitting problem and to allow the proposed system to generalize on the test dataset and to give better results. Python 3.6.8 is used for implementing the proposed system. Experiments are done on CPU with 32 GB of RAM, Intel processor with NVIDIA GPU and Ubuntu operating system.

Performance Evaluation Measures

Performance of the proposed system is calculated using several measures as described in the following:

(i) Confusion matrix: Table 1 presents classification algorithm performance.

(ii) Classification Accuracy Rate (CAR): CAR measures basing on confusion matrix as given in the following equation [22]:

$$\mathrm{Accuracy} = (tp + tn)/(tp + tn + fp + fn) \tag{1}$$

(iii) Precision: Measures relevancy of the result generated as defined in the following equation [22]:

$$\mathrm{Precision} = tp/(tp + fp) \tag{2}$$

(iv) Recall: Measures relevancy of the result generated as given in the following equation [22]:

$$\mathrm{Recall} = tp/(tp + fn) \tag{3}$$

(v) *F*-Measure as given in the following equation [22]:

Fig. 8 Visualizing the 68 facial landmark coordinates[16]

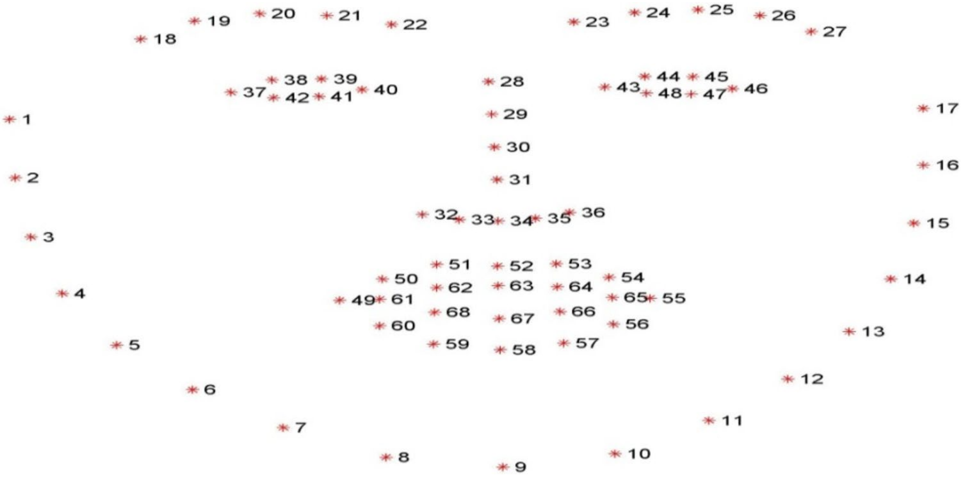


Fig. 9 Sample input dataset

Fig. 10 Sample output with age prediction

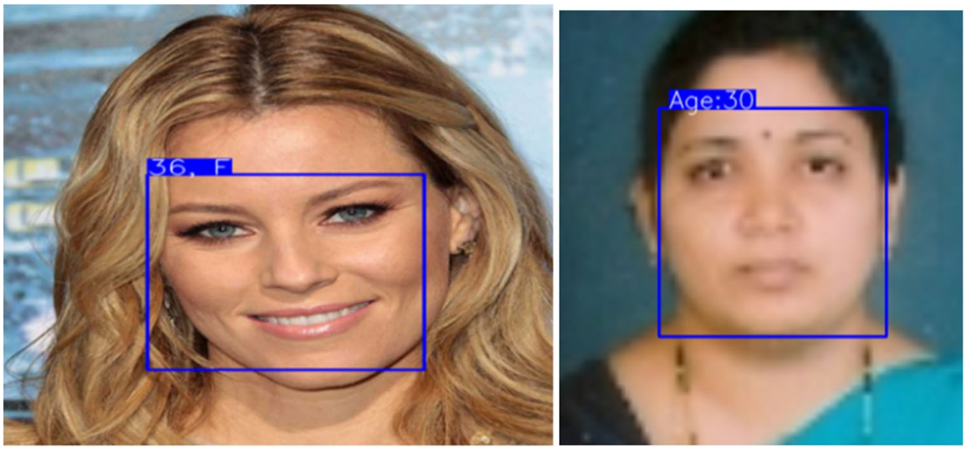


Table 1 Confusion matrix [22]

	Predicted $a=0$	Predicted $b=1$
actual $a=0$	TP	FP
actual $b=1$	FN	TN

Table 2 Results of the proposed system for predicting age

S. No	Dataset	No of faces	No of faces correctly classified
1	APPA real dataset	7470	7153
2	College dataset	40	37

Table 3 Accuracy of the proposed system

S.No	Dataset	Precision	Recall	*F* Measure	Accuracy
1	APPA real dataset	98.9	95.9	0.97	95
2	College dataset	93.9	96.8	0.95	93

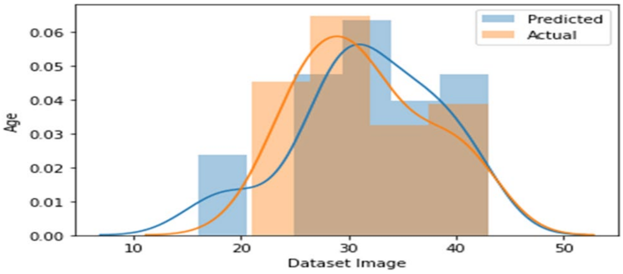


Fig. 11 Distribution graph for actual vs predicted age

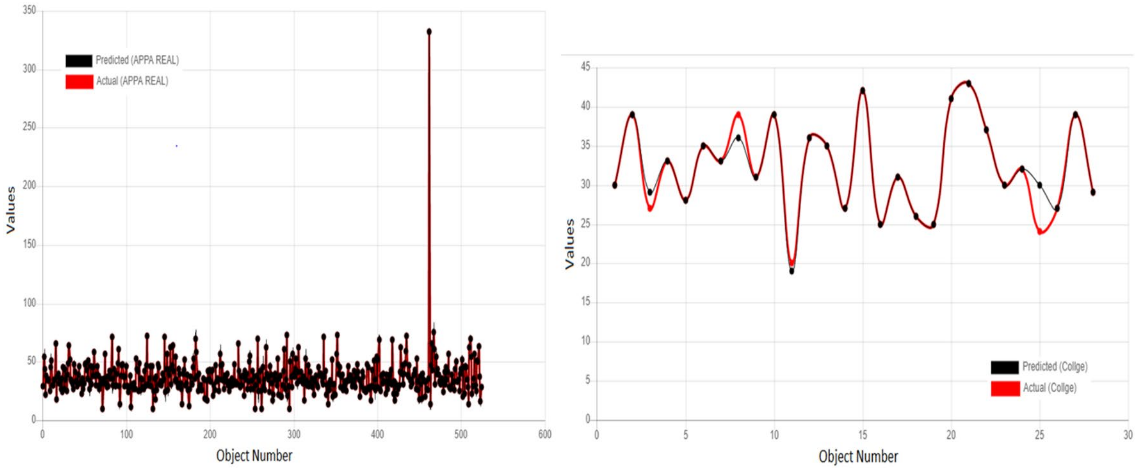
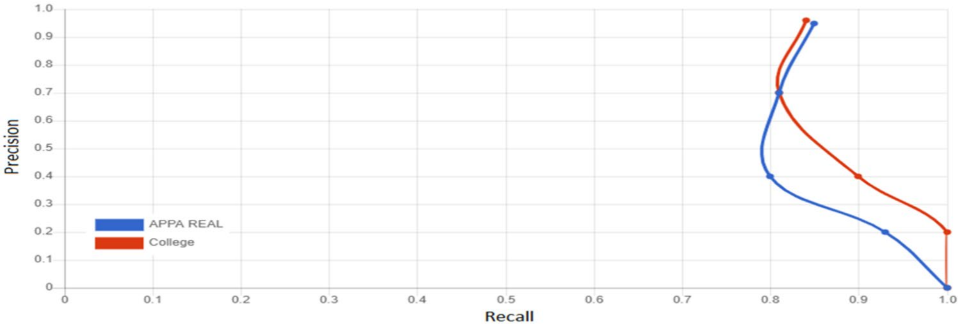


Fig. 12 **a** Correlation between actual and predicted. **b** Correlation between actual and predicted

Fig. 13 Precision recall curve



$$\mathrm{F1} = 2 \times (\mathrm{precision} \times \mathrm{recall})/(\mathrm{precision} + \mathrm{recall}) \tag{4}$$

(vi) Cross-entropy loss function H is given in Eq. (5) [23]. This value can be between 0 and 1. The low this value, more robust the developed model.

$$H(p, q) = -\sum_{x} p(x) \log q(x), \tag{5}$$

where $p(x)$ is the actual probability and $q(x)$ is the predicted probability.

(vii) Mean Absolute Error (MAE) is given in the following equation:

$$\frac{1}{n} \sum_{i=1}^{n} (xi - x) \tag{6}$$

Results and Analysis

Initially accuracy is calculated to evaluate the proposed model. Other methods such F1 score is calculated to find balance of precision and recall.

Table 2 presents results of the proposed system on various datasets for predicting age. It can be observed that

Table 4 Mean absolute error (MAE)

S.No	Work	MAE
1	[24]	4.08
2	[25]	5.98
3	[26]	3.688
4	[27]	4.4
5	[28]	1.65
6	Proposed system	1.61 for APPA Real 1.54 for College

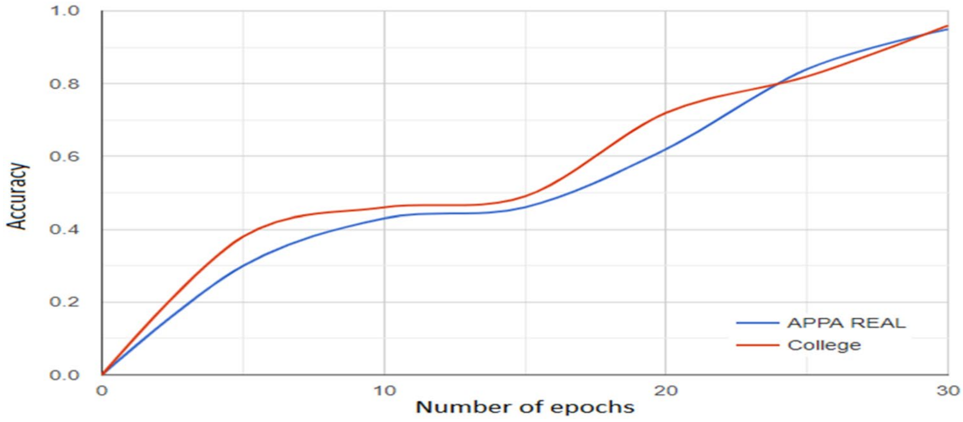


Fig. 14 Number of epochs vs accuracy

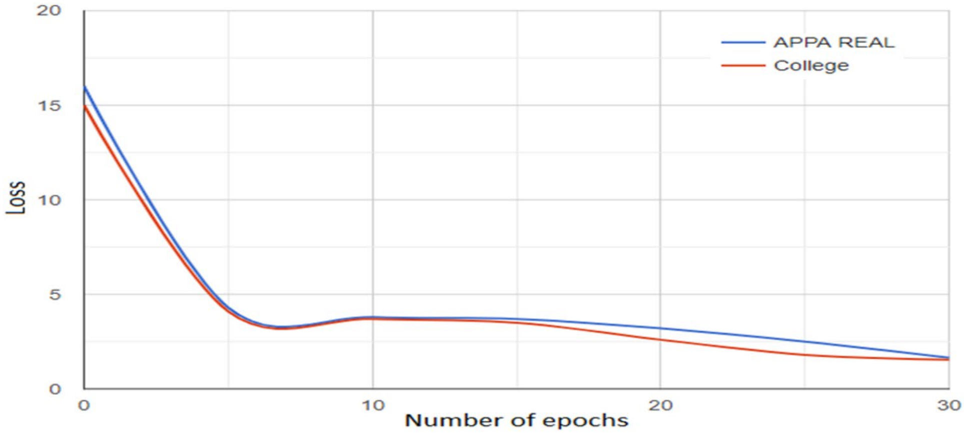


Fig. 15 Number of epochs Vs loss

false positives are less, because Wide ResNet predicts object boundaries after thorough scan on the entire image. Few objects are not detected because of cluttered background area. IoU of 0.3 also helped in reducing false positives. It was observed that few objects were not detected as shown in these tables because of the following reasons:

(i) Occlusions.
(ii) The basic nature of Wide ResNet on the overlapping objects and merged them as a single object.
(iii) Setting threshold of 0.5 to discard weak detections (partial objects are not detected).

Initially accuracy is calculated to evaluate the proposed model. Other methods such as Precision and Recall are used to evaluate the performance of proposed system. Also, F1 score is calculated to find balance of precision and recall. Table 3 presents the accuracy of the proposed system for predicting age in the two datasets.

It was observed that prediction is correctly done for faces with semi-closed eyes or images with single face. Accuracy for the age bins 0–10 and 50–70 showed 3% less accuracy. This is because of closed eyes, wrinkles are not properly segmented and few are presented with more age. Accuracy of prediction is nearly 97% for the bins 10–50 and 70–100.

Figure 11 presents distribution graph for predicted and actual age in the dataset. The *X*-axis represents the dataset image and the *Y*-axis represents the age. The predicted age is represented in blue color and the actual age is represented in cream color.

Figure 12a and b presents correlation between predicted age and actual age on validation dataset. Figure 13 presents Precision recall curve for the two datasets.

Table 4 presents mean absolute error (MAE) of various works to predict age.

Figure 14 presents number of epochs vs accuracy and Fig. 15 presents number of epochs vs loss. Learning rate is 0.1 and number of epochs are 30. Most ResNet models were close to convergence after around 20 epochs and here they are stable when nearing to 30. It can be observed that accuracy increases and loss decreases as the epochs reach 30 resulting to 95% accuracy for APPA real dataset and 96% for college dataset. While loss decreased to 1.61 for APPA REAL and 1.54 for college dataset.

Conclusions and Future Work

Deep residual networks can scale to thousands of layers to improve performance of given task, but face diminishing feature and reuse problem. ResNet blocks can be used to grow width of the network and decreases network depth, called wide residual networks (WRNs). To predict the age, we apply a deep convolution neural network to the face image detected.

Initially facial key points are extracted; subsequently SART is used for sampling and finally predicts one of the 101 classes. The proposed system is tested on two datasets and the results found to be accurate. The proposed system is compared with existing works and results proved that it is better than existing works. MAE and correlation coefficient are also evaluated. Our future work is to improve the proposed system for image objects with less illumination and occluded objects. Our work also concentrates to predict age of moving objects in a video surveillance where automatic feature extraction is a challenging task.

Funding Not applicable.

Declarations

Conflict of interest Authors have no conflicts of interest.

Ethics approval Ethics approval was not required for this study.

References

1. Quividi, https://quividi.com/content-intelligence/. Last Accessed 15 Apr 2019
2. Just how precise is Quividi's gender classification?, https://quividi.com/how-precise-is-quividis-gender-classification/. Last Accessed 15 Apr 2019
3. Chauhan NS (2019) Predict age and gender using convolutional neural network and openCV, https://towardsdatascience.com/predict-age-and-gender-using-convolutional-neural-network-and-opencv-fd90390e3ce6, Last Accessed 2 Jan 2019
4. Ramos-Muguerza E, Docío-Fernández L, Alba-Castro JL (2018) From hard to soft biometrics through DNN transfer learning, 2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS), 2018, pp 1–7
5. Sun Y, Lo FP, Lo B (2019) A deep learning approach on gender and age recognition using a single inertial sensor. In: 2019 IEEE 16th international conference on wearable and implantable body sensor networks (BSN), 2019, pp 1–4
6. Neal TJ, Woodard DL. You are not acting like yourself: a study on soft biometric classification, person identification, and mobile device use. IEEE Trans Biometr, Behav, Identity Sci. 2019;1(2):109–22.
7. Jiang M, Guo G. Body weight analysis from human body images. Trans Inf Forensics Secur. 2019;99:1–1.
8. Levi G, Hassner T (2015) Age and gender classification using convolutional neural networks. In: IEEE workshop on analysis and modeling of faces and gestures, 2015, p 1–9
9. Agbo-Ajala O, Viriri S. Face-based age and gender classification using deep learning model. In: Dabrowski J, Rahman A, Paul M, editors. Image and video technology. PSIVT 2019. Lecture notes in computer science, vol. 11994. Cham: Springer; 2019.
10. Nwankpa C, Ijomah W, Gachagan A, Marshall S (2018) Activation functions: comparison of trends in practice and research for deep learning. arXiv Prepr. arXiv1811.03378
11. Srivastava RK, Greff K, Schmidhuber J (2015) Training very deep networks, in advances in neural information processing systems
12. Zagoruyko S, Komodakis N (2016) Wide residual networks. In: Proceedings Br. Mach. Vis. Conf. 2016, May 2016, p. 87.1–87.12
13. APPA-REAL (real and apparent age). http://chalearnlap.cvc.uab.es/dataset/26/description/ Last Accessed 2 Jan 2019
14. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) Mixup: beyond empirical risk minimization. In arXiv: 1710.09412
15. https://github.com/yu4u/mixup-generator. Last Accessed 2 May 2019
16. Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on computer vision and pattern recognition, 2014, p 1867–1874
17. Brown A, Mettes P, Worring M (2019) 4-Connected shift residual networks, ICCV Workshop 2019, p 1–8
18. https://www.dlology.com/blog/easy-real-time-gender-age-prediction-from-webcam-video-with-keras/. Last Accessed 2 Jan 2019
19. Bobrov E, Georgievskaya A, Kiselev K, Sevastopolsky A, Zhavoronkov A, Gurov S, Rudakov K, Tobar MD, Jaspers S, Clemann S. PhotoAgeClock: deep learning algorithms for development of non-invasive visual biomarkers of aging. Aging. 2018;10(11):3249–59. https://doi.org/10.18632/aging.101629.
20. Biamby G, Fair D, Karapetov A, Karapetov B (2018) Age and gender classification using deep neural networks. https://medium.com/@andreykar_79244/age-and-gender-classification-using-deep-neural-networks-a8ded298a838. Last Accessed 2 Jan 2019
21. Andersen AH, Kak AC. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. Ultrason Imaging. 1984;6(1):81–94.
22. Vasavi S, Shaik R, Yarlagadda S. Moving object classification in a video sequence using invariant feature extraction. Hershey: IGI Global; 2018. p. 1–25.
23. Murphy K. Machine learning: a probabilistic perspective. Cambridge: MIT; 2012.
24. Agustsson E, Timofte R, Escalera S, Baro X, Guyon I, Rothe R (2017) Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database. In: IEEE International conference on automatic face and gesture recognition (FG), 2017, p 1–8
25. Clapes A, Bilici O, Temirova D, Avots E, Anbarjafari G, Escalera S (2019) From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation CVPR conference, p 2486–2495

26. Rondeau J, Alvarez M (2018) Deep modeling of human age guesses for apparent age estimation. In: International joint conference on neural networks (IJCNN), Rio de Janeiro, 2018, pp 01–08
27. Gunjal A, Abin D. Survey on age estimation system. Int J Comput Appl. 2018;182(3):1–4.
28. Debgupta R, Chaudhuri BB, Tripathy BK. A wide resnet-based approach for age and gender estimation in face images. International Conference on innovative computing and communications. Advances in intelligent systems and Computing, vol. 1087. Singapore: Springer; 2020.
29. Deng Y, Luo P, Loy CC, Tang X (2014) Pedestrian attribute recognition at far distance. In: 22nd International conference on multimedia

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.