© April 2020 | IJIRT | Volume 6 Issue 11 | ISSN: 2349-6002

Sex Determination from Ear Images using Deep Learning

Srishti Kohli[1], Rahul Bajaj[2]
[1]*Indira Gandhi Delhi Technical University for Women, Delhi*
[2]*Vellore Institute of Technology, Chennai*

Abstract- **Similar to fingerprints, facial Recognition and iris scan, ear can also act as a unique identifier for a person. Hence, Otomorphology has been a hot research topic lately. We present a complete analysis of gender classification based on ear images using custom convolutional neural networks and Transfer learning using pre-trained deep convolutional neural networks.**
Similar Research has been performed on ear images to classify gender but this exploration is over dataset with images captured in unconstrained conditions.

I.INTRODUCTION

Otomorphology, also known as earology is the study of the physical features of the ear, which like fingerprints shows a unique pattern for every individual. Owing to its position on the face, the ear is highly suitable for conduction of passive biometrics wherein authentication is done in absence of the subject's awareness of it. Prerequisite for this passive biometrics is an uncontrolled environment. Hence, our research incorporates exploration over a dataset consisting of images of ear captured in an unconstrained environment. There are limitations to all the biometrics Techniques. Face Recognition is tough when facial expressions are altered, on application of cosmetics on it, or identical twins. Similarly, fingerprints of senior citizens, Cancer patients are difficult to capture.
Further benefits of Ear Recognition are in the field of Forensics - Body Identification. The common techniques used for identifying human remains include Facial Recognition, Fingerprints analysis, DNA analysis etc. However, in possible situations like distortion of face or loss of fingerprints in Cancer patients, these methods are likely to fail. As an alternative to these, Ear recognition may be used. Studies suggest that the structure of the ear does not change drastically with time, making it an excellent source of data to act as an identifier.

This paper intends to determine sex/gender using ear images. Our Research focuses on leveraging the high performance computing power to contribute to this interdisciplinary field of Computer Vision.

II. DATA

The data used is the fourth version of dataset EarVN1.0, released in 2020. The EarVN1.0 dataset was curated by collecting images of the ears of 164 Asians in the year 2018. It comprises 28,412 coloured ear images captured in an unconstrained environment. This is followed by cropping ears from the images of faces captured over a large variation of scale, illumination and pose. The dataset provides images of both ears per person. Out of 164 people, 98 are males and 66 are females. Sex Determination via ear images can be performed and evaluated on this dataset since it provides 17,571 ear images of male and 10,841 images of female.
We apply various deep learning algorithms on this dataset to accomplish sex determination from ear images.

Fig I:(a) Sample data

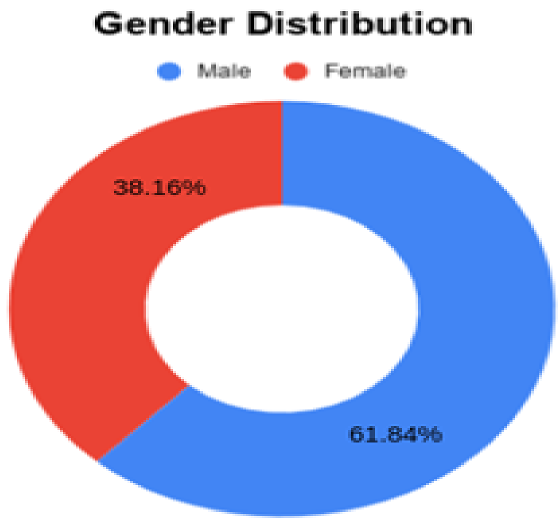

Fig I: (b) Male and Female Distribution in Dataset

III. ALGORITHMS

Computer Vision is a branch of Computer Science which incorporates the complex biological functioning of sight to a computer system. It is about processing digital images, videos etc. making visual perception understandable for machines. Intuitively, Mathematics describes convolution as a mathematical operation on two functions to generate a third function which identifies cross-correlation between them.

A. Convolutional Neural Network

Convolutional Neural Network comprises a network of nodes/neurons, to learn the weights and biases, employing convolution as the linear operation. It can be understood as replacing matrix multiplication in a neural network with convolution in at least one of its layers. The neural network receives pixel points as the input. The input layer provides the attribute values (pixel points) to the subsequent layers. Each node in the subsequent layer learns the weights based on the input from the previous layer and gives the output as the activated result of the learned weights. The network tries to extract features of the image frame to develop embeddings for the same. These embeddings are then pushed to a classifier to assign them the target class labels.

The Custom CNN used in this research uses 80:20 train-test data. 64X64X3 (RGB) images were given as the input. The first layer after the input layer(first hidden layer) consists of 32 filters each of size 3 X 3 X 3, keeping stride as 1 and padding the same. We get 28 parameters per filter. The next layer is Max Pooling with kernel size 3 giving a convolved image of size 21 X 21 X 3. The next layer, Convolution layer consists of 64 filters each of size 3 X 3 X 3, keeping stride as 1 and padding the same. This is followed by Max Pooling with kernel size 3 giving a convolved image of 7 X 7 X 3. The succeeding layer is again, Convolution layer with 64 filters, each of size 3 X 3 X 3, keeping stride as 1 and padding the same. Max Pooling layer with kernel size 3 giving convolved output of size 2 X 2 X 3 follows. The next layer consists of 128 filters, each of size 3 X 3 X 3, keeping stride as 1 and padding the same. The flattened output of 512 dimensions is fed into the next layer i.e. Dense layer having output dimension as 128. The final layer is Dense Layer having output dimension 1.

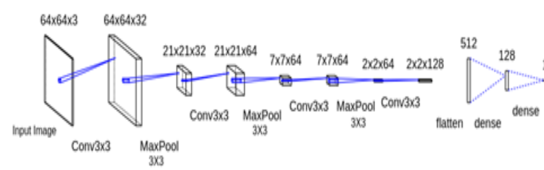

Fig II: Custom CNN layers

B. ResNet-50

ResNet-50 is a pre-trained convolutional neural network with a depth of 50 layers and 25,636,712 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by a 50 layer deep network, ResNet-50. Since we use ResNet for feature extraction, the last layer(classifier) of the network is excluded. This network takes 64 X 64 X 3 images as input and yields a flattened output of length 2048. This is fed to our 8 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,128,64,64,32,1.

C. VGG19

VGG19 is a pre-trained convolutional neural network with a depth of 26 layers and 143,667,240 parameters. This deep learning model is used for image classification and has weights trained on

ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by a 26 layer deep network, VGG19. Since we use VGG19 for feature extraction, the last layer(classifier) of the network is excluded. This network takes 64 X 64 X 3 images as input and yields a flattened output of length 512. This is fed to our 8 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,128,64,64,32,1.

D. Xception

Xception is a pre-trained convolutional neural network with a depth of 126 layers and 22,910,480 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by a 126 layer deep network, Xception. Since we use Xception for feature extraction, the last layer(classifier) of the network is excluded. This network takes 64 X 64 X 3 images as input and yields a flattened output of length 2048. This is fed to our 8 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,128,64,64,32,1.

E. NASNetLarge

NASNetLarge is a pre-trained convolutional neural network with a depth of 1210 layers and 88,949,818 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by a 1210 layers deep network, NASNetLarge. Since we use this for feature extraction, the last layer(classifier) of the network is excluded. This network takes 64 X 64 X 3 images as input and yields a flattened output of length 4032. This is fed to our 8 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,128,64,64,32,1.

F. InceptionV3

InceptionV3 is a pre-trained convolutional neural network with a depth of 159 layers and 23,851,784 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by a 159 layers deep network, InceptionV3. Since we use this for feature extraction, the last layer(classifier) of the network is excluded. This network takes 75 X 75 X 3 images as input and yields a flattened output of length 2048. This is fed to our 8 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,128,64,64,32,1.

G. DenseNet201

DenseNet201 is a pre-trained convolutional neural network with a depth of 201 layers and 20,242,984 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by a 201 layers deep network, DenseNet201. Since we use this for feature extraction, the last layer(classifier) of the network is excluded. This network takes 64 X 64 X 3 images as input and yields a flattened output of length 1920. This is fed to our 8 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,128,64,64,32,1.

H. InceptionResNetV2

InceptionResNetV2 is a pre-trained convolutional neural network with a depth of 572 layers and 55,873,736 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by a 572 layers deep network, InceptionResnetV2. Since we use this for feature extraction, the last layer(classifier) of the network is excluded. This network takes 75 X 75 X 3 images as input and yields a flattened output of length 1536. This is fed to our 8 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,128,64,64,32,1.

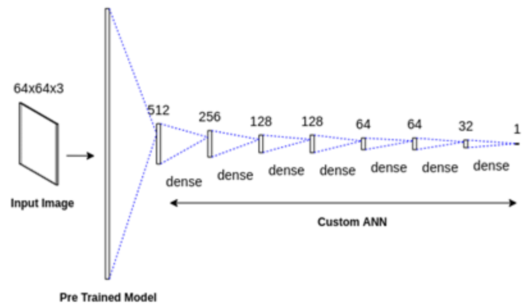

Fig III: Transfer Learning

III. RESULTS

The Sex Determination was performed using a couple of different approaches. This includes training over the dataset to generate domain specific embeddings. The Custom Convolutional Neural Network with 9 layers generated 128 embeddings and an accuracy of 92.92% and 195,969 parameters. All training and testing is done on NVidia K80 16GB GPU.

Model	No. of Layers	Accuracy (%)
Custom CNN	9	92.92
ResNet50	57	83.02
InceptionV3	166	71.09
VGG19	33	81.29
DenseNet201	208	77.97
NASNetLarge	1217	68.95
Xception	133	63.60
InceptionResNetV2	579	69.87

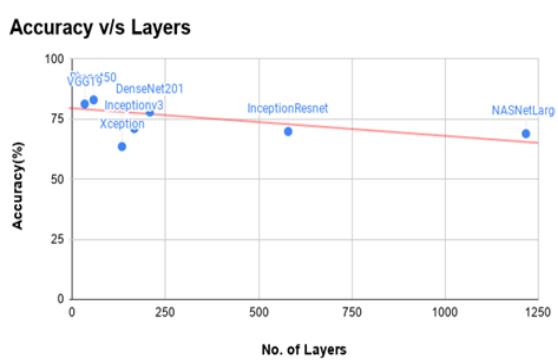

Fig IV: Accuracy v/s No. of Layers

Model	No. of Epochs	No. of Steps	Time per Epoch(s)	Accuracy (%)
Custom CNN	20	700	95	92.92
ResNet50	10	500	81	83.02
InceptionV3	10	353	34	71.09
VGG19	25	353	32	81.29
DenseNet201	10	353	43	77.97
NASNetLarge	10	353	44	68.95
Xception	10	353	33	63.60
InceptionResNetV2	10	353	44	69.87

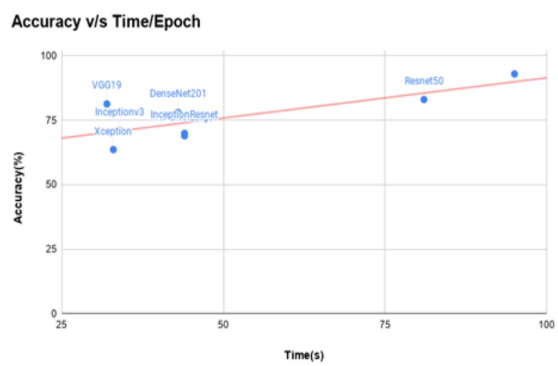

Fig V: Accuracy v/s Time per Epoch

State of the art Pre-trained models were used for implementing transfer learning. Embeddings obtained using the non-trainable layers were passed to a standardized dense layer network for classification. ResNet50, InceptionV3, VGG19, DenseNet201, NASNetLarge, Xception, InceptionResNetV2 produced accuracy of 83.02%, 71.09%, 81.29%, 77.97%, 68.95%, 63.60%, 69.87% respectively

Model	No. of Parameters	Accuracy (%)
Custom CNN	195,969	92.92
ResNet50	24,832,065	83.02
InceptionV3	23,047,137	71.09
VGG19	20,482,305	81.29
DenseNet201	19,500,801	77.97
NASNetLarge	87,176,979	68.95

Xception	22,105,833	63.60
InceptionResNetV2	55,318,945	69.87

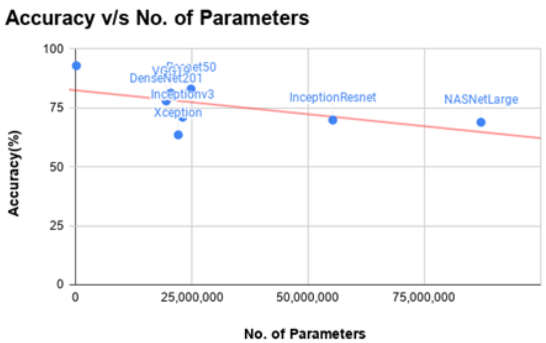

Fig VI: Accuracy v/s No. of Parameters

IV. CONCLUSION

Custom Convolutional Network trained specifically for this dataset resulted in the maximum accuracy of 92.92%. This model showed significantly more accurate results as compared to the pre-trained model which is in sync with our intuition.

All the transfer learning techniques portrayed a linear trend (approximately) for Accuracy v/s Time/Epoch. ResNet50 was the most efficient.

Most pre-trained models showed similar values for Accuracy v/s No. of parameters apart from NASNet Large and InceptionResNetV2.

We conclude a successful exploration with a possibility of future research in person identification with the images of ear forming the basis of more robust methods of Authentication and Biometrics.

REFERENCES

[1] Truong Hoang, Vinh (2020), "EarVN1.0", Mendeley Data
[2] arXiv:1512.03385
[3] arXiv:1409.1556
[4] arXiv:1512.00567
[5] arXiv:1602.07261
[6] arXiv:1610.02357
[7] arXiv:1608.06993
[8] arXiv:1707.07012
[9] chollet 2015 keras, title=Keras, author=Chollet, Fraçois and others, year - 2015