Age, Gender, Emotion detection using Machine Learning

By Mahavir Chandaliya
San Jose State University, San Jose, CA.
mahavir.chandaliya@sjsu.edu

***Abstract*— In this paper, we have developed an application of neural networks and computer vision to predict age, gender and emotion of the person captured in the video from digital devices. In the times we live in, there has been a tremendous development in information technology, every person owns or uses a computer or a mobile device. Things like virtual meetings, video calls are very common. Social media has seen a phenomenal growth in today's world. "Age and gender identification have become a major part of the network, security and care", it has applications in age specific content control, layered advertisements and marketing as well as security to a large extent [1].**

In this research paper, we have used Machine learning techniques for extracting facial data and predicting the age, gender and emotions or sentiment of the person captured in the video feed. We have used face image datasets and facial expressions datasets for training the models developed in this research. We have also utilized computer vision technology for capturing the facial data required for our prediction. Right from enriching the customer experience of their products to implementing targeted advertisements and marketing companies today are greatly dependent on Artificial intelligence technologies and data analytics. "The customer demographics such as age and gender, the sentiment a customer experiences towards particular products plays a significant role in the sales and operations of retailers and small scale vendors" [3].

Recognizing age, gender as well as emotions of a person is also highly useful in various other sectors such as psychology, security as well as mental health. The effectiveness of the proposed solution is evaluated on real-life data using live video feed from digital device of the user.

Keywords—Age; Gender; Emotion; Object detection; Machine learning; Computer Vision; Neural Networks; OpenCV; MTCNN; Convolution Neural Network; Facial Recognition; Deep Learning; Numpy;

1. INTRODUCTION

Recently, the world went through the pandemic of coronavirus also known as Covid-19; a communicable disease that spreads very fast from close contact with the person tested positive for Covid-19. Preventing the spread of this disease has been one of the most important things in these difficult times. Virtual meetings on different online platforms such as Google meetings zoom and Microsoft teams are being used by a large number of people and institutions. Live streams and use of webcams have become largely used medium for organizing virtual events and interactions. This massive increase in use of webcams and video cameras have solved a lot of problems that we have faced due to this dangerous contagious disease but have also given a rich source of facial data that can be further explored. In this paper, such an application is developed that recognizes a person's age, gender and emotion from the facial data obtained from the images through webcam and video cameras which can be utilized in various fields like business, psychology etc.

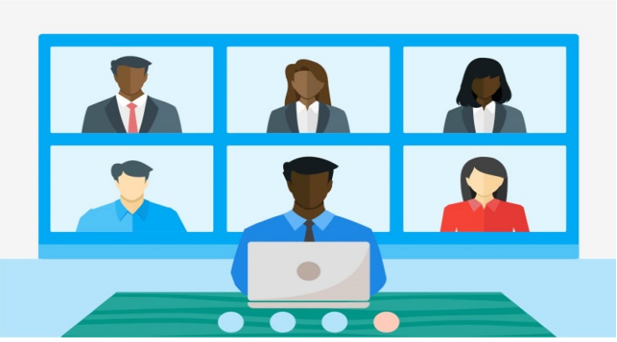

Fig. 1: A Virtual Meeting online

The application developed in this research work detects a person's age, gender and emotion based on the facial data obtained from the user's webcam. The application uses the programming language python and different sets of libraries to implement the given functionalities. "OpenCV" is a python library to manipulate image and videos and to read images and videos from the hardware. We have used MTCNN python module for face detection from the captured video. MTCNN stands for 'Multi-task Cascaded Convolutional Networks', which is a framework built using python libraries of OpenCV and PyTorch that detects faces from an image. Different features of the face are recognized such as eyes, nose and mouth to detect faces from an image frame of the video.

The core of the application uses three models to detect age, gender and emotion of a person developed using neural networks, these models are trained on datasets to predict the output from the facial data. The models consist of various layers that read the input data from an image datasets, after training on datasets with output labels the models correctly predict the age, gender and emotion of the person. For training the models, we have used various open source datasets available on the web.

The emotion model developed in the research is trained on the CKPlus Facial Emotion dataset obtained from the website ‘kaggle’ which houses a large number of open source datasets for machine learning. The CKPlus facial Emotion dataset contains 981 images with labeled emotions. The emotion labels comprise of ‘anger, contempt, disgust, fear, happy, sadness and surprise’.

The Age and Gender models are trained on the UTKface datasets which contains 20k images which are labeled with person’s age, gender and race. The UTKface dataset is available for public on github and is an excellent source of facial images data to train the machine learning models.

Even though the study uses Machine Learning libraries of python for implementation of face detection, we have used a deep learning algorithm for predicting age, gender and emotion values. The main difference between deep learning and machine learning is that, deep learning carries out the task of feature extraction and classification on its own, one only need to provide a pre-processed labeled dataset for training. In contrast, using when using a simple machine learning model one needs to provide the model with features required for prediction. Deep learning is a subset of Machine learning.

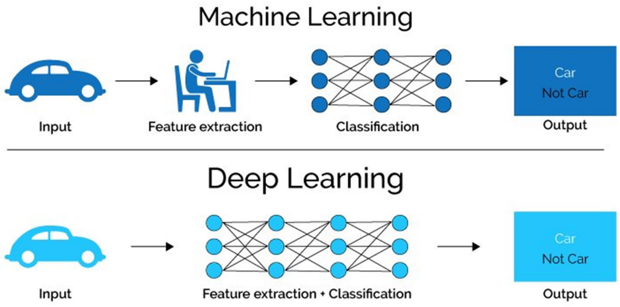

Fig. 2: Face deformation due to direct re-sizing

Firstly, when we talk about object detection using deep learning we need to understand what a convolution neural network is. A Convolution Neural Network is a deep learning algorithm that take an input of an image assigns weights and biases to various segments of an image and learns to differentiate one image from another. Here, the images with age, gender and emotion labels are used as input to CNN which learns from the images, calculates appropriate weights and biases to develop a model capable of predicting the values of these labels for a new image from the knowledge gained by training on the known dataset.

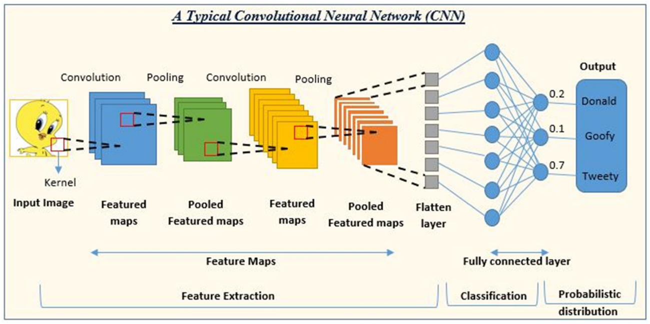

Fig. 3: A Convolution Neural Network

The CNN algorithm is comprised of various layers that extracts features from an image and learns from these features and the labels of known images. In the fig. 3, a CNN implementation is depicted where various layers of the neural network can be seen. Each layer consists of nodes which contain the weights assigned to various features obtained from an image. In our case, facial features such as eyes, nose and mouth show different variations from person to person based on emotion, age and gender. CNN learns from these labeled image datasets and once the model is developed it looks for similar trends in the facial data of a new image and predicts the values of age, gender and emotion.

The Age, Gender and Emotion of a person can be further supplied to Artificial Intelligence systems to give further analysis of the demographics. Businesses can use this data to develop better customer relationships, increase sale of their products by targeting relevant markets and simply understand their customers better. “The customer demographics such as age and gender, in addition to the sentiment a customer experiences towards particular products plays a significant role in the sales and operations of retailers and small scale vendors” [2].

Real time age, gender and emotion predictions can also be utilized in the field of security and protection. Monitoring a person’s emotional state and being equipped with age and gender estimations of that individual could allow the security companies to more conveniently track down people and face unknown security threats.

Educational institutions can conduct surveys and obtain feedback from these predictions obtained from students during online lectures, which can be helpful to further develop the structure of the classes and educational materials that can enhance the educational experience of the students and further improve the standards of the institution. The results of this research can be applied to many data analytical systems to assist with their operations.

2. LITERATURE SURVEY

In this paper, we have developed a methodology for detecting Age, Gender and Emotion using a person's facial data acquired from a webcam or a video device. A similar implementation can be seen in the study, "Age and Gender Prediction using Deep Convolutional Neural Networks" [1] where a CNN model is developed for predicting age and gender of an individual using Haar Cascade face classifier which even though detects the faces accurately but is more simpler and slower compared to other face detection algorithms. We have used MTCNN for face detection which provides better performance compared to Haar cascade classifier and can be easily implemented in python. MTCNN is also easily integrated with other python libraries making it easy for quick prototyping. Another research work [8], also uses Haar Cascade to detect faces and predict age and gender.

Another notable research in this field is detection of age, gender and emotion from voice [2]. In this research study, the authors have implemented a machine learning application that predicts the age, gender and emotion values from recorded audio of a person's voice. The only disadvantage of this research is the difficulty in implementing the application in real time as the audio generally is filled with noise and overlapping sounds from various sources. Compared to audio, use of a video makes it easy to detect objects and differentiating individuals from one another in the image. However, in the future once more technologies are developed in the field of speech recognition, a combined application that uses both images and audio captured from a video can lead to a more accurate and highly reliable solution.

Using an image gives us access to more amount of data that can be used for analysis, "An image is worth a thousand words" [6]. Machine Learning is considered a subset of Artificial Intelligence (AI). Using Machine Learning techniques, the machine makes intelligent decisions as well as predicts future outcomes based on the patterns in the data without explicit planning. In the study, "Weakly Supervised Emotion Intensity Prediction for Recognition of Emotions in Images" [4] the prediction is divided in three streams, a classification stream, an emotion intensity prediction stream and another classification stream. This methodology of predicting emotions from emotional intensity maps generated from an image gives a more accurate emotion prediction, however the increased complexity of the system downgrades its performance making it unsuitable for real time use. To overcome this issue instead of generating an output in three different stages, we have used a CNN model with more number of layers that increases accuracy while giving a high performance. Increasing the number of layers, even though increases the complexity of the model it allows us to easily optimize the neural network as we have to work on a single model.

"The main challenge of visual emotion recognition is that emotions are much higher levels of abstraction than visual semantics" [4]. Predicting emotions can never be completely accurate however; with development of new technologies and more amounts of data in the future this can be achieved.

Age estimation from facial data from images plays an important role in forensics or social media [9]. Age, gender and emotion prediction also has wide applications in the field of security and crime sector. Use of facial data is widely being used for the purpose of marketing. "The knowledge of the average shopper in addition to the expressions one portrays while shopping, plays a paramount role" [3]. In the research [3], the customer sentiments are portrayed as, overall satisfaction, loyalty and future engagement. Analyzing this the facial prediction and combining them with this metrics can help to develop marketing strategies that can lead to expanding customer base and considerably increase the profitability of the business.

In the research named "Criminality from Face" [7], using deep learning and facial data the study predicts whether the person is criminal or not. This study uses a similar CNN model for implementation. The model developed in this research can provide a similar application in the field of crime. Knowing a person's age and gender not only helps in the criminal investigation but with emotion prediction one can arrive at a more accurate conclusion regarding the crime committed.

Another research study [5], takes into account various factors such as illumination, color or RGB values in an image to further develop a more diverse and efficient CNN model for facial data detection. Using the findings of this research, our study can be further improved to perform in diverse environments with various type of camera equipment. A different work [10] has differentiated numerous methods with a detailed study of their impacts, and based on these findings we have improved the CNN model used in this research. By using more number of epochs to train the model, the model accuracy is seen to improve.

Many other research in this field uses technologies which require a very high computing power to perform real time analysis and to train the deep learning algorithm which increases its cost and maintenance Proposed system uses Python OpenCV and machine learning libraries which can run easily on any Linux or windows machine and thus can be more easily and widely used. Thus, finding a greater reach in implementing this system in various sectors.

Emerging technologies can provide great help to implement computer vision applications and with the help of new technologies in the field of robotics the study developed in this research can be applied in the sector of surveillance and other fields of national security. Drones and robots equipped with camera and an intelligent system can gather a lot more information from dangerous places such as warzones compared to a human being without any risk of life on such missions.

3. IMPLEMENTATION

The application is implemented in python using neural networks and OpenCV where a video stream is given as input and it outputs a video with rectangular boxes around each person's face highlighting the boxes with age, gender and emotion predicted values. The steps for implementing the application can be seen in Fig. 4. In summary, OpenCV captures the video from the camera or a webcam. Every frame captured from the video is converted to RGB format and sent to MTCNN functions, where the faces are detected in the image frame and for every face the trained models predict the values of age, gender and emotion which are rendered back on the video frame with a bounding box to generate the final output.

The age model is more complex compared to the gender and emotion model developed in the research. Each model is comprised of various CNN layers which are convolution layers, dense layers, pooling layers, an input and an output layer. Each layer uses an activation function of 'relu' which assigns random weights and biases to the nodes of the neural network before they are computed when the model is trained. The models are compiled with Adam optimizer function that is needed for developing the gradient or fitting the curve based on the data to predict new values. Each model uses a different loss function based on the dataset and labels. We have used 'mse' for age model. It stands for Mean square error which is used as the output label is in numerical format. For gender model, we have used 'binary_crossentropy' loss function as the output is binary i.e. either male or female, and for emotion model as there are 7 categories of emotions as out put we have used the loss function of 'categorical_crossentropy'.

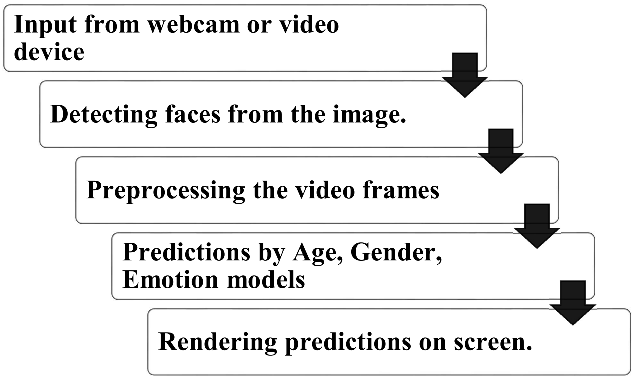
Input from webcam or video device
Detecting faces from the image.
Preprocessing the video frames
Predictions by Age, Gender, Emotion models
Rendering predictions on screen.

Fig. 4: Process flow of the implementation

Firstly before obtaining the video input from the camera and applying the models, we trained the models for age, gender and emotion on the training datasets. The datasets that we used for training emotion model is called CKPlus Facial Emotion dataset obtained from kaggle which has grey scale images with a fixed size of 48*48 and for age and gender models we used the UTKface datasets. The datasets are used to train the models so that we can further use the trained models to predict the age, gender and emotion values from the image frame obtained from the video device.

Datasets supplied to train the models are needed to be pre-processed as the model only requires the specific labels and the facial image of the person from the images dataset with a standard size. The dataset used for emotion detection i.e. the CKPlus Facial Emotion dataset contains the images which are already cropped and aligned allowing us to quickly train the model on these images without any need for pre-processing.

The UTKface dataset contains images which need to be resized for training the model as deep learning models require images to be in standard sizes. Direct resizing is an easy way to achieve this. However, it results in image deformation. Deformed images affect the model's performance leading to inaccurate results.

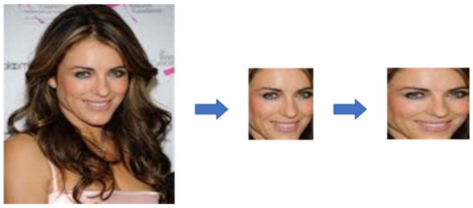

Fig. 5: Face deformation due to direct re-sizing

We have used centered resizing which is ideal for ensuring accurate training of the models. It keeps the face in the center of the image with required size and no distortion. The input images are required to be of the size (224, 224, 3) for both models.

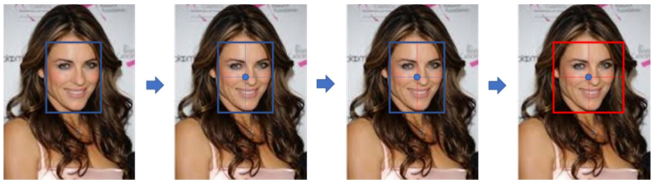

Fig. 6: Centered re-sizing

Now the datasets are ready to for us to start training our models. We have used various python libraries such as numpy and keras for computations and designing the models. Python supplies us with built-in methods that we can use for implementing various layers of our deep learning models. One can easily optimize the models for better performance by fine-tuning the function parameters. The models are than trained on the datasets and saved as a pickle file which is another API that python provides for saving trained models and use them wherever required. This functionality is provided by Python pickle library. Once we have our trained models, now we have to supply appropriate input of face images from the webcam or the video device to carry out the predictions.

For detecting faces we have used the algorithm called MTCNN which stands for Multi-task Cascaded Convolutional Networks. MTCNN is used for face detection and alignment from an image which consists of three steps which make the use of Convolution neural network to detect faces and recognize face features such as eyes, nose and mouth. When implementing MTCNN using python, we are equipped with a pre-trained model which is optimized allowing us to use the model directly.

MTCNN returns a list of python dictionaries where each dictionary represents the faces detected in an image. The dictionary consists of a 'box'; x and y coordinates of the bounding box surrounding the face with width and height. Secondly, 'key-points' which represents the facial features such as nose, ears and eyes. And lastly 'confidence' which comprises of the model's score of confidence for the faces detected or measure of surety of whether the object detected is a face or not. The score varies between 0 and 1, 1 being the highest.

Once the faces are detected, we supply the image frames to the trained models for predicting age, gender and emotion. The system gives an output comprising of age, gender and emotion of the individual detected in the image frames from the video. For the output we have used, Open CV for putting the age, gender and emotion value on the image along with a bounding box highlighting the person's face.

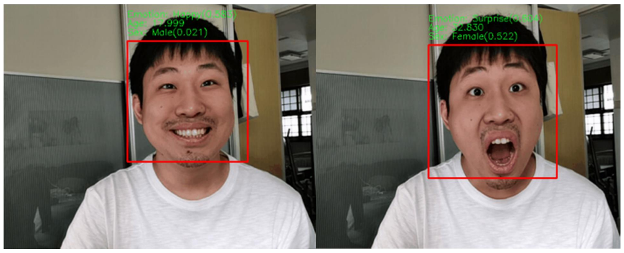

Fig. 7: Illustration of output of the application

4. RESULTS AND FURTHER IMPROVEMENTS

We used a video from the Webcam of the computer device to test out the developed application. We found out that the algorithm detects the faces accurately and prints out the age, gender and emotion values on the image. The gender values are seen quite more accurate compared to the age and emotion predictions. The predictions verified manually as the video cannot contain depict the emotion of a person. The manual verification of the results suggests that the output obtained is considerably accurate.

Due to constraints on computational resources we have used the UTKFace dataset which has 5K images of faces for age and gender emotion, model performance can be improved by using a larger dataset. Similarly, the better emotion dataset can improve the performance of the emotion model and also give results that comprises of more number of emotions than the basic 7 emotion labels that we have used in our prediction.

The devised methodology in this paper can be further improved to provide a real time analysis of facial data by using a faster model which further provides insights into various other aspects of the person. Along with displaying the age, gender and emotion of the person the application can keep the track of the emotional state of a person recorded from the video device and give an output with a more detailed analysis giving deeper insights.

Occlusion is a challenge when developing object detection applications in Computer Vision. Occlusion occurs when we want to see something, but are unable to due to some event or the way our sensor is setup. This generally occurs if an object we are trying to locate is hidden (occluded) by some other object. To overcome this problem and to improve the methodology a model which is more resistant to occlusions can be used.

Another challenge faced in developing computer vision systems is Automatic calibration. The process of determining internal camera parameters directly from a number of un-calibrated images of unstructured scenes is Auto Calibration [13]. Implementing Automation can result in further improvement of this methodology making it more effective. Furthermore, by using a NVIDIA GPU or a High power GPU, this system can run in real-time.

The predictions obtained can be used to develop business solutions and also can be supplied to more complex Machine Learning application for further analysis.

5. CONCLUSION

In this paper a deep learning application capable of predicting an individual's age, gender and emotion from a video device or a webcam is developed. The proposed application uses the deep learning algorithm of CNN to carry out these predictions.

In these times, providing virtual solutions for business meetings and public events has been one of the most important challenges in front of various technological companies. The devised system works on top of these solutions to provide further insights into the data obtained from these virtual spaces which makes it equally useful and easy to implement.

Applications of this research in the field of business, security and education sectors are a topic of further research.

6. REFRENCES

[1] I. Rafique, A. Hamid, S. Naseer, M. Asad, M. Awais and T. Yasir, "Age and Gender Prediction using Deep Convolutional Neural Networks," 2019 International Conference on Innovative Computing (ICIC), 2019, pp. 1-6, doi: 10.1109/ICIC48496.2019.8966704.

[2] S. R. Zaman, D. Sadekeen, M. A. Alfaz and R. Shahriyar, "One Source to Detect them All: Gender, Age, and Emotion Detection from Voice," 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 2021, pp. 338-343, doi: 10.1109/COMPSAC51774.2021.00055.

[3] E. P. Ijjina, G. Kanahasabai and A. S. Joshi, "Deep Learning based approach to detect Customer Age, Gender and Expression in Surveillance Video," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225459.

[4] H. Zhang and M. Xu, "Weakly Supervised Emotion Intensity Prediction for Recognition of Emotions in Images," in IEEE Transactions on Multimedia, vol. 23, pp. 2033-2044, 2021, doi: 10.1109/TMM.2020.3007352.

[5] K. Jhang and J. Cho, "CNN Training for Face Photo based Gender and Age Group Prediction with Camera," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2019, pp. 548-551, doi: 10.1109/ICAIIC.2019.8669039.

[6] K. Mohan, A. Seal, O. Krejcar and A. Yazidi, "Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-12, 2021, Art no. 5003512, doi: 10.1109/TIM.2020.3031835.

[7] K. W. Bowyer, M. C. King, W. J. Scheirer and K. Vangara, "The "Criminality From Face" Illusion," in IEEE Transactions on Technology and Society, vol. 1, no. 4, pp. 175-183, Dec. 2020, doi: 10.1109/TTS.2020.3032321.

[8] Harshitha H S, Soujanya CK, Suhas S K, Swathi S Gowda, Dr. Shashikala S V, "Age Gender And Emotion Identification Using Face Recognition," 2021 International Research Journal of Modernization in Engineering Technology and Science Volume:03, e-ISSN: 2582-5208.

[9] Sidharth Nair, Dipesh Nair, Gautam Nair, Anoop Pillai and Prof Sujith Tilak, "Detection of Gender, Age and Emotion of a Human Image using Facial Features," 2020 International Research Journal of Engineering and Technology Volume: 07, e-ISSN: 2395-0056 p-ISSN: 2395-0072.

[10] Manasa S.B., Jeffy S. Abraham, Anjali Sharma, Himapoornashree K.S., "Age, Gender and Emotion Detection using CNN," 2020 International Journal of Advance Research in Computer Science Vol. 11, ISSN No. 0976-55697.

[11] H. M. Abdel Mageed and A. M. El-Rifaie, "Automatic calibration system for electrical sourcing and measuring instruments," *2013 12th International Conference on Environment and Electrical Engineering*, Wroclaw, 2013, pp. 30-34, doi: 10.1109/EEEIC.2013.6549578.