International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 297-303
Published Online May 2020 in IJEAST (http://www.ijeast.com)

AGE, ETHNICITY AND EMOTION RECOGNITION USING MTCNN

Anuj Mishra, Amit Kumar Pandey, Ahmad Affan, Akarsh Kumar
Department of CSE
IMS Engineering College, Ghaziabad, U.P., India

Dr. Swati Singh
Department of CSE
IMS Engineering College, Ghaziabad, U.P., India

***Abstract*— Age, ethnicity, and emotion are quite important attributes of people that determine their interests. Knowledge of age and ethnicity has also been found to improve medical diagnosis as different diseases are more prevalent in different races and age groups. Most of the existing works are focused on identifying age, ethnicity, and emotion in images having a single person, and their faces well-lit and facing forward are nearly trivial, however, real applications need to consider different face angles and differences in lighting conditions.**

Work by Sneha Thakur [1] et. al. uses Artificial Neural Network to identify the age of a person but requires faces to front-facing. Other such works such as by S. Hma Salah, et. al. [2] and Xiaoguang Lu et. al. [3] identifies ethnicity using Haar Wavelet Features and Linear Discriminant Analysis but still have similar image restrictions. Also, these methods only focus on cases where only one person is present per image but realistically many people could be present in a single image.

Identifying people's interests and giving proper product recommendations is a hot field in consumer analytics. A proper identification model that works on realistic images can give a boost to this domain. In this paper, we are going to propose a method to identify age, ethnicity, and emotion in images having many people facing different ways and in different lighting conditions. Our approach firstly extracts faces using a pre-trained face localization model and then feeding them to separate neural network models, one each for detecting age, ethnicity, and emotion.

Keywords— **age, emotion, ethnicity, Multi Task Cascade Neural Network (MTCNN)**

I. INTRODUCTION

The human face is most highly rich stimulus that provides diverse information for adaptive social interaction with people. Humans can process a face in a variety of ways to categorize it by its identity, along with a number of other demographic characteristics, including ethnicity (or race), emotion and age. These tasks are quite trivial for humans. A lot of effort has been devoted in the biological, psychological, and cognitive sciences areas, to discover how the human brain perceives, represents, and remembers faces. However, imparting these abilities to a machine has been at the forefront of research in the last few years.

Human behavior and social interaction vary widely on the basis of geography, age group, peer group, emotional state and ethnicity. Various habits are found to transition from good to bad and bad to good between races, religion and geographical regions. Consumer analytics and product recommendation engines can benefit a lot if it can automatically infer these characteristics of its users. It can uncover a large positive pattern and improve profits to a great extent.

Various works have also proposed that considering human demographic characteristics for medical diagnosis have shown improved result providing better accuracy. These methods require explicit inclusion of demographic statistics into the data. The process can be streamlined if it can automatically infer these characteristics automatically.

Existing methods such as those by S. Hma Salah, et. al. [2] and Xiaoguang Lu et. al. [3] used Haar Wavelet Features and Linear Discriminant Analysis respectively to identify ethnicity. Work by Sneha Thakur et. al. is a seminal work that identifies the age group of people using artificial neural network. However, these methods have few common drawbacks. First, they strictly require the faces to be front facing rather than sideways. Second, they can process images which have only one face.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 297-303
Published Online May 2020 in IJEAST (http://www.ijeast.com)

Ilias Maglogiannis et. al. [4] proposed a method to identify emotions by monitoring eyes and mouth expression of people. But this method also has the earlier stated drawbacks.

In this paper, we have proposed a model that overcomes these widely ignored drawbacks and can identify multiple faces in an image, along with their demographic statistics such as emotion, age group and ethnicity. We have used the DataTurks dataset containing 120 images mapped with ethnicity, emotion and age of different people present in the image. Our approach firstly converts the image to grayscale, and then uses a pre-trained MTCNN [5] model to extract faces from the images. These grayscale faces are fed to different deep learning models, one each for identifying age, emotion and ethnicity. We were able to reach 58% accuracy, hugely limited by the small size of the dataset used.

II. EXISTING WORK

A. Age Detection

Both regression and classification problems can be used for the study of age classification where a certain group is associated with a certain age and the age of the person in the image can be classified in one of the groups.

At the very first stage of face recognition, the facial features are extracted from the face, at this extraction stage, many different techniques have been used in the past. Kwon et al [6] used anthropometric models for the first time for the extraction of the facial age feature and to divide ages into categories: infant, youth, elderly based on the craniofacial development theory and skin wrinkle features.

Another method of the learning process as introduced by Guo et al. [7] for the estimation of the age of faces in the image where the face images are transformed into low-dimensional manifolds i.e. high dimensional face datasets are mapped into low-dimensional manifolds. This manifold learning was very helpful in reducing the amount of training data for the classification tool.

Recently in the past few years, deep learning technologies like CNN are used for the estimation of age and achieved better outcomes in age detection. Dong et al.[8] was the first one to use CNN for age estimation and was successful in designing a complete age estimation system.

Wang et al. [9] propose a method for the age estimation where CNN id only used to extract the features of the face and then the result obtained in used as an input for a classification or regression model which is then used for the estimation of the age, this makes the feature avoids the limitations of the hand-designed features.

B. Ethnicity Identification

A large number of algorithms have been developed in the past for the detection of the ethnicity of faces in the image.

Hosoi et el. [10] proposed an approach with the 3 categories: Asian, European, and African. He integrated the Gabor wavelet and retina sampling in his work. This approach proposed by them was used with the support vector machine (SVM). Although the accuracy for each ethnicity Asian, African, and European: 96%, 94%, and 93% respectively but this approach has issues when other ethnicities are to be detected.

Viola-Jones [11] provided efficient work for the detection of the ethnicity of the faces with varying conditions like camera variation, face pose, the color of an image; etc. They use the AdaBoost classifier for detection. This approach was very fast on the datasets with the above conditions in the images.

S. M. M. Roomi used the Viola-Jones [11] algorithm for ethnicity detection. Various features of the face are extracted from the image after the detection of the face. The dataset used in this classification problem is Yale, FERET dataset. This approach has an accuracy of 81%.

C. Gender Identification

Just like age and ethnicity, there are many research works that are already been done and presented for gender detection from a face in the image.

Geetha et al. [12] proposed a new gender detection algorithm where the author uses two databases one is the FEI database and the other is a self-built database and tests are performed in these databases. From the face in the image, different types of texture features were extracted and these were from several levels that are global directional and regional. For the classification stage of this proposed approach, a kernel-based SVM as used.

Toews et al. [13] proposed a hybrid model for the detection of gender and age of the person in the image using a combined framework. Eidinger et al. [14] also proposed a hybrid model in which the author proposed the use of a hybrid pip-line for age and gender-related study.

In gender detection also Deep Convolution Neural Network was effective with its outstanding performance for the image recognition feature. Levi et al. [15] applied CNN based methods to extract the features of the images as well as the classification algorithms for gender detection.

Duan et al. [16] proposed a hybrid approach for gender and age detection. The author of the paper proposed the use of CNN for the extraction of the features and using an Extreme Learning Machine (ELM) for the classification. ELM-CNN is the name given to this model by the author. MORPH-II and Adience are the two public databases on which the ELM-CNN was evaluated.

III. PROPOSED MTCNN

We propose using MTCNN with joint dynamic weight loss to classify gender, age and race and further mitigate related

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 297-303
Published Online May 2020 in IJEAST (http://www.ijeast.com)

biasness. The proposed method utilizes disjoint features of the fully connected layers of a Deep CNN employing separated fully connected layers for fulfilling this multi-task learning that operates to aim for better face attribute analysis. It exploits the synergy and therefore the disjoint features among the tasks, boosting up performances. We exploit the very fact that information contained in CNN the features are hierarchically distributed throughout the network. Lower layers contains feature like edges and corners, and thus contain better localization features.

Hence they're more suitable for learning localization and pose estimation tasks. Whereas, on the opposite hand, deeper layers, e.g., higher top layers are class-specific and suitable for learning complex tasks like face recognition and therefore the fully connected layers involve for the classification task i.e., where the top to finish system can learn and plan to discriminate the salient features for various inherent tasks during a MTCNN scenario. Given these aforementioned MTCNN-characteristic, also because the aim of the work to reinforce face attribute analysis, we propose to customize Facenet for the face recognition with ResNet V1 inception (as it's one among the prominent face architectures). This Facenet network consists batch of input layer and a Deep CNN architecture (ResNet V1 in our scenario) followed by L2 normalization, which results in face embedding. This is often followed by the triplet loss during training. The architecture consists of a stream of convolution layers, normalization layer and pooling layers will be followed by 3 inception blocks and their reduction. The latter followed by dropout and a totally connected layer. Now, we split the network into three separate branches like the various classification tasks (i.e., gender, race and age). We add three fully connected layers, one for race classification, gender classification and age classification. Finally, a Softmax layer is added to every of the branches to predict the individual task labels feature with L2 normalization and respective dropout layer. After each convolution a Rectified linear measure (ReLU) is deployed as activation function. The Facenet model turns a picture of a face into a vector of 128 floating point numbers. These 128 embedding are often used as features for classification.

While using the Facenet we fine-tuned the fully connected layer. In the MTCNN the network is split into fully connected layer followed by individual SoftMax layer of every task. Therefore an input layer tuple of the MTCNN, for the given training set T with N images contains T = Ii, Yi, where i=1: N, where Ii is that the image and Yi may be a vector consisting of the labels. In MTCNNs it's challenging to define the loss weight for every task. In previous works, this was dealt either by treating all tasks equally Dong et al, dynamic MTCNN Fang et al., [17], obtaining weights via brute-force search or by dynamically assigning disjoint weights for the side task. However neither of these strategies adds our setting. Unlike pose and illumination, gender, race and age classification are closely related countenance. Moreover, they possess varying degrees of relevance for both the intra-class and inter-class variation. Such quality depends on their intensity exhibited per samples.

Hence, we seek to optimize the effect of this multi-task facial attribute classification (i.e., gender, age and race) by learning them jointly and dynamically, counting on the degree of relevance of the feature present for every classification task. Specifically, the MTCNN should directly learn classification of task relations from data rather than subjective task grouping, thereby deciding weight of the task sharing. Hence, we propose joint dynamic weighting scheme to automatically assign the loss weights for the each task during training.

First, we discover the summed weight for the each classification task by brute force search on the validation set. Further by adding a totally connected layer and a SoftMax layer to every task the model gets proficient to shared features from the last common layer, which is aimed toward learning the dynamic weights (for each iteration) depending on degree of relevance of the task. Therefore we obtain the dynamic weight percentages for every task from the fully connected layer. Further, the function of the soft-max layer converts the dynamic weights to positive values that sum to 1. Consequently, the foremost relevant task is to contribute predominantly to the ultimate loss and therefore the additional task is to contribute to the relevant task, so as to scale back the loss of the foremost relevant task. Thereby, the MTCNN should assign a better weight for a non-relevant task with a lower loss, in order to scale back the general loss. A mini-batch Stochastic Gradient Descent (SGD) was employed to unravel the above optimization problem of loss weight. Further, the weights are averaged for every batch.

NEURAL NETWORK ARCHITECTURE

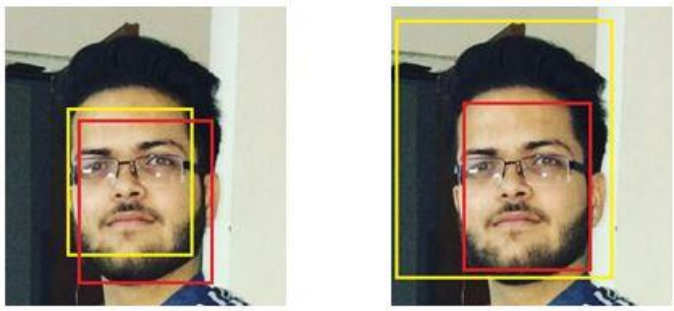

Figure 1: Example of large and small overlap

The neural network architecture that is being used in this model is Multi-task Cascaded Convolutional Network (MTCNN). This model has three convolutional networks (P-Net, R-Net, and O-Net).

1. Stage 1:

In order to detect faces of different sizes, we have to create an image pyramid. Or we can say that we want to create different

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 297-303
Published Online May 2020 in IJEAST (http://www.ijeast.com)

copies of the same image in different sizes to search for different sized faces within the image.
For each image we have a kernel/filter that will scan the image starting from top left corner. The portion of image which has been scanned is passed through P-net which returns the coordinates of a bounding box if it notices a face. Then it would repeat the same procedure for other section of the image. Number of pixels that the kernel moves every time is called a stride. We have used a stride of 2 as having a stride of 2 helps reduce computation complexity without significantly sacrificing accuracy.

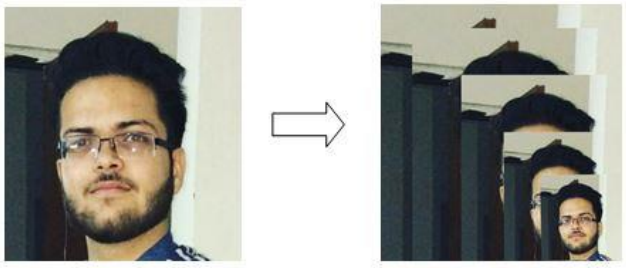

Figure 2: Example of conversion of pyramid image from original image

The only disadvantage is that we have to recalculate all indexes related to the stride. For example, if the kernel detected a face after moving one step to the right, the output index would tell us the top left corner of that kernel is at (1,0). However, because the stride is 2, we have to multiply the index by 2 to get the correct coordinate: (2, 0).

After passing in the image, we need to create multiple scaled copies of the image and pass it into the first neural net — P Net — and gather its output**.** The network is more confident about some boxes compared to others.

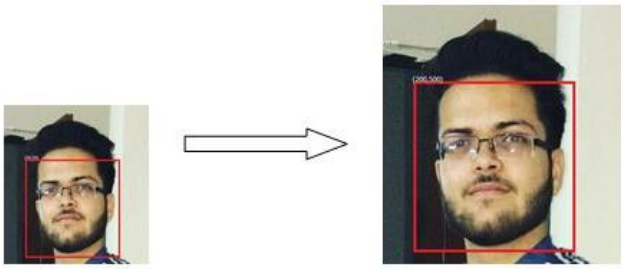

Figure 3: Example of conversion of smaller scaled image to larger unscaled image

Thus, we need to parse the P-Net output to get a list of confidence levels for each bounding box, and delete the boxes with lower confidence (i.e. the boxes that the network isn't quite sure contains a face). However, there are still a lot of bounding boxes left, and a lot of them overlap. **Non-Maximum Suppression**, or NMS, is a method that reduces the number of bounding boxes. Afterward, we convert the bounding box coordinates to coordinates of the actual image.

2. Stage 2:

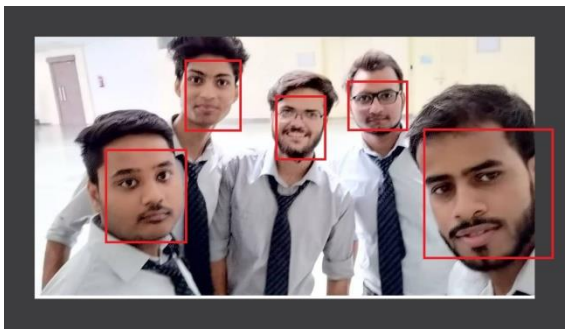

Figure 4: Example of bounding box in an image having multiple numbers of people

Sometimes, it may possible that an image may contain only a part of a face peeking in from the side of the frame. For every bounding box, we create an array of the same size, and copy the pixel values (the image in the bounding box) to the new array. If the bounding box is out of bounds, we only copy the portion of the image in the bounding box to the new array and fill in everything else with a 0. This process is called padding. After padding we resize them and normalize them between -1 to 1.

Now that we have so many image arrays, we will feed them into R-net and gather its output. R-Net's output is similar to that of P-Net: It includes the coordinates of the new, more accurate bounding boxes, as well as the confidence level of each of these bounding boxes. Once again, we get rid of the boxes with lower confidence, and perform NMS on every box to further eliminate redundant boxes. Since the coordinates of these new bounding boxes are based on the P-Net bounding boxes, we need to convert them to the standard coordinates. And after that we pass the image through to O-net.

3. Stage 3:

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 297-303
Published Online May 2020 in IJEAST (http://www.ijeast.com)

The outputs of O-Net are slightly different from that of P-Net and R-Net. O-Net provides 3 outputs: the coordinates of the bounding box, the coordinates of the 5 facial landmarks, and the confidence level of each box. Once again, we get rid of the boxes with lower confidence levels, and standardize both the bounding box coordinates and the facial landmark coordinates. Finally, we run them through the last NMS. At this point, there should only be one bounding box for every face in the image.

IV. EXPERIMENT AND RESULT

The Python library of Facenet1 is employed to calculate facial embedding of face images and developing the proposed MTCNN. The Facenet library was implemented in TensorFlow. It includes pre-trained Facenet models for face recognition. The models are validated on the LFW database and were trained on a subset of MSCeleb-1M database. The models architecture follows the Inception-ResNet-v1 network. Facenet library includes the implementation of the detection, alignment and landmark estimation, as proposed by Zhang et al, which we use for preprocessing for our images. The output of our proposed MTCNN may be a 128 dimension floating point embedding, almost like Facenet. The Scikit-learns SVM version with RBF kernel is employed with Tensorflow for classification of this embedding for face recognition because the pre-trained model was trained on a way larger face dataset but with less similarity in reference to face attributes, we employ transfer learning. Specifically we freeze all the initial layers of the pre-trained model and train the last top layers. Therefore, the highest layers (which are known to contain the face attribute information) are customized to our interest. During transfer learning we make sure that the ultimate layers aren't restored from the pre-trained model and that we also to make sure that gradients are gated for all other parameters during training. While fine tuning the load, decay is about 0.0005.

V. CONCLUSION

Various studies have been performed till date on face recognition and using facial features for different operations. In this paper we have discussed the trends that already exist, take a look over the challenges associated with them, and proposed using MTCNN architecture to identify faces, along with convolutional neural networks to identify gender, age and ethnicity of individuals. The convolutional neural network are being used to detect the faces and determine the age, gender and ethnicity of all the faces in the image at the same time. The person can either be facing directly into the camera or facing sideways. We have achieved acceptable results which can be further improved by using wider dataset of images of multiple persons in different orientations. The neural networks used here can be further optimized using more thorough hyper-parameter optimization. We hope that the methods discussed here will serve as a base for those who want to contribute to the growing field of multiple face-recognition at the same time.

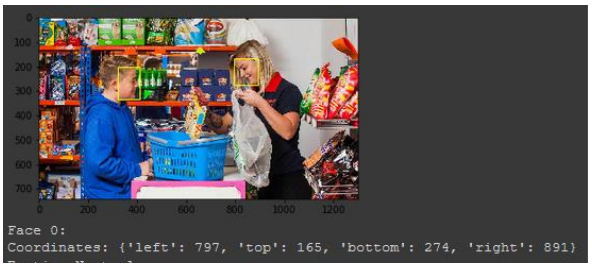

Figure 5: Image showing result after applying it on an image

VI. ACKNOWLEDGEMENTS

Our Sincere thanks to guide Mrs Swati Singh (Professor in IMS Engineering College, Ghaziabad) for her patient guidance and constructive suggestions for the research in the area of Age, Ethnicity and Emotion Detection using MTCNN. This research is made possible due to her willingness to devout her time generously.

VII. REFERENCES

[1] Thakur Sneha, and Verma Ligendra. (2012). Age identification of Facial Images using Neural Network, in International Journal of Computer Science and Information Technologies, (pp.4244 – 4247).

[2] Hma Salah S., Du H., and Al-Jawad N. (2013). Fusing Local Binary Patterns with Wavelet Features for Ethnicity Identification. International Journal of Computer, Information, Systems and Control Engineering, (pp.347 - 353).

[3] Lu Xiaoguang, and K. Jain Anil. (2004). Ethnicity Identification from Face Images. In the proceedings of 2004 conference on Biometric Technology for Human Identification, (pp.114 - 124).

[4] Maglogiannis Ilias, Vouyioukas Demosthenes, and Aggelopoulos Chris. (2009). Face detection and recognition of natural human emotion using Markov random fields. Personal and Ubiquitous Computing, (pp.95 - 101).

[5] Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, and Qiao Yu. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Processing Letters, (pp.1499 – 1503).

[6] H. Kwon Young, and da Vitoria Lobo Niels. (1999). Age Classification from Facial Images. Computer Vision and Image Understanding, (pp.1 – 21).

[7] Guo Guodong, Fu Yun, R. Dyer Charles, and S. Huang Thomas. (2008). Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust

Regression. IEEE Transactions on Image Processing, (pp.1178 – 1188).

[8] Yi Dong, Lei Zhen, and Z. Li Stan. (2014). Age Estimation by Multi-scale Convolutional Network. Asian Conference on Computer Vision, (pp.144 - 158).

[9] Wang Xiaolong, Guo Rui, and Kambhamettu Chandra. (2015). Deeply-Learned Feature for Age Estimation. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, (pp.534–541).

[10] Hosoi S., Takikawa E., and Kawade M. (2004). Ethnicity estimation with facial images. In the Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition, (pp.195 - 200).

[11] Viola P., and Jones M. (2004). Robust Real-Time Face Detection. In Proceedings of International Journal of Computer Vision, (pp.137–154).

[12] Geetha A., Sundaram M., Vijayakumari B. (2019). Gender classification from face images by mixing the classifier outcome of prime, distinct descriptors. Soft Computing, (pp.2525–2535).

[13] M Toews, and T Arbel. (2009). Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, (pp.1567 – 1581).

[14] Eidinger Eran, Enbar Roee, and Hassner Tal. (2014). Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, (pp.2170 – 2179).

[15] Levi Gil, and Hassncer Tal. (2015). Age and gender classification using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, (pp.34 – 42).

[16] Duan Mingxing, Li Kenli, Yang Canqun, and Li Keqin. (2018). A hybrid deep learning CNN–ELM for age and gender classification. Neurocomputing, (pp.448 – 461).

[17] Fang Yuchun, Ma Zhengyan, Zhang Zhaoxiang, Zhang Xu-Yao, and Bai Xiang. (2017). Dynamic multi-task learning with convolutional neural network. In the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, (pp.1668 - 1674).