Automatic Face Image Labelling of Ethnicity, Emotion, Age, and Gender

Juli Kyada[1], Harsh Kakadiya[1], Khushi Donga[1], and Nishant Koshti[1]

Department of Artificial Intelligence and Machine Learning
CSPIT, CHARUSAT, Gujarat, India
julikyada293@gmail.com

Abstract. To build the accurate and strong computer vision models, big and trustworthy data sets are necessary. Conventionally, manual facial image annotation, which includes age, gender, emotion, ethnicity, etc., has been carried out. Manual annotation is good but time-consuming, expensive, and labor-intensive and is particularly vulnerable to inconsistencies due to human bias and subjective judgments, which can have adverse impacts on the quality of the dataset and model.
In a bid to solve these problems, the current paper suggests an Automatic Facial Image Labeling System that is built upon the semi-supervised learning methods by using pre-trained models. The suggested solution will allow automatic and correct labelling of facial pictures and greatly increase the speed of generating the dataset, enhancing consistency and reliability. The system reduces errors in annotation and improves the level of scalability because of the human intervention being minimized.
The suggested framework will enable generating large, diverse, and high-quality annotated datasets within a reduced time interval, which will find applications in practical purposes in healthcare, human-computer interaction, as well as other areas of computer vision. The paper is a contribution to the more scalable, more efficient, and reliable processes of data annotation in the contemporary computer vision studies.
The proposed model performs well in age and gender classification (89), but is still difficult in ethnicity (63) and emotion (68). These findings highlight the complexity of multi-attribute facial recognition and demonstrate the potential of semi-supervised learning for reducing the need for labeling

Keywords: Automatic Facial Image Labeling · Semi-supervised Learning · CNN · FixMatch · Facial Attribute Recognition

1 Introduction

The pillars of machine recognition with machines have been machine vision and AI-based applications through which a range of social networking to virtual user support and security surveillance has been done. Although it is possible to find many systems that are constructed in a manner that they perceive individual cues like age, gender or emotion and few systems perceive a compounding of low

features like: race - emotions - age - sex at a time. Multi-feature simultaneous face recognition is a very difficult thing to do relying on the delicacy of the human appearance, biases on the data set and failure to produce a single prediction model.

The bias in the data set is one of the largest setbacks of the facial analysis. The majority of the publicly accessible databases of facial images are skewed to the objects of the white race to the extent that they can result in worse performance of the system by other ethnic groups, which restricts the generalization of AI systems to other groups.Several studies have highlighted the impact of demographic bias on facial recognition systems and emphasized the need for more balanced datasets and fairness-aware evaluation [1,11,8].

The solution to this problem is that FairFace will provide over 100,000 pictures of the same amount of seven ethnicities with the age and gender data, and, consequently, makes the process of model evaluation more equitable besides annotating the facial images with the help of AI and neural networks feasible and efficient than before possible as well [7]. The semi-automated approaches, such as interactive labeling system by Tian et al., employ an unsupervised clustering with the minimum annotations and hence demand significantly less annotation effort and are also as precise to a large extent 20. Almost in the same manner, Zhang and other individuals developed a dynamic annotation system on the probabilistic networks that involves automated prediction of such annotations with minimum human intervention that contributes to speeding up and improving the quality of labeling operations thereof in essence[18].

Big face databases are also required to produce sound models. Anvari and Athitsos developed an entire system of automatically generating the information of the faces through the use of unlabeled images in order to reduce the load of the manual generation of data sets, thus establishing a fully automated system of generating face data information [2]. Besides, GAN-generated artificially datasets have also been found to be helpful as an alternative in model training and assessment. Colbois and others demonstrated that StyleGAN2 artificial identities can be built at scale with no difficulty, and can give the same assessment scores as real datasets, and, because that is unacceptable in law, and can be built at scale and with the same assessment scale as real datasets, which is also important since that is ambiguous law-wise and can be safely built at scale, giving the same evaluation scores as real datasets, which is an essential fact because that is intolerable in law[4].

Background Research The identification of facial characteristics has evolved very fast because of the invention of deep learning. The first experiments that focused on predicting age with VGG-16 architecture (DEX) were the experiments on predicting age with CNN on Adience databa by Rothe (2016); Levi and Hassner also experimented with predicting age and sex on Adience databa with CNN [9]. In terms of emotional identification, AffectNet provided deep models with large quantities of labeled variables in natural settings to be conditioned on natural settings [12].

Although the developments have been made, most of the research studies on these characteristics concentrate on either one or two of these traits. The phenomenon of ethnic identification has not been studied adequately and in part because of ethical issues and imbalanced data sets. It was addressed by FairFace by generating an ethnically balanced list so as to be able to come up with more equitable and generalizable models. The multiplicity of tools and techniques such as Tian and cite have also been used in the creation of quality datasets that facilitate the multi-attribute facial recognition systems, including [15,18].

2 Related Work

Deep learning has shown great progress in facial attribute recognition. Some of the pioneering works, such as the DEX framework, used pre-trained VGG-16 model for age estimation, showcasing the potential of deep convolutional networks for facial recognition tasks.Advanced loss functions and embedding methods such as SphereFace, CosFace, and ArcFace have significantly improved discriminative facial feature learning [10,16,5]. Likewise, Levi and Hassner introduced CNN-based methods for age and gender classification on the Adience dataset [9]. In the context of emotion recognition, large-scale datasets like FER2013 and AffectNet have enabled the training of deep learning models that can better cope with real-world facial expressions, datasets such as VGGFace2, MS-Celeb-1M, and CelebV-HQ have played an important role in improving the robustness and diversity of face analysis models [3,**?**,19]. Moreover, new datasets such as FairFace dataset and UTKFace offer more balanced labels for age, gender, and ethnicity to avoid gender and racial biases in facial recognition systems.

Various techniques have been proposed to minimise manual annotation of facial datasets. Conventional approaches are heavily dependent on manual annotations, which are costly. Semi-automated methods based on clustering or unsupervised learning have been proposed to alleviate this problem [15,18,2]. Such techniques cluster facial images and label them with little human intervention, thus facilitating annotation.

however, clustering methods can lead to unreliable labels due to unsupervised nature. Furthermore, such methods are typically not suitable for multi-attribute prediction, and tuning the parameters is crucial for their performance [15,18]. Scalability is also often an issue when applying these approaches to large-scale multi-attribute data sets [2].

However, while these methods have made progress, most of them primarily focus on single or bi-attribute prediction, with little research into scalable multi-attribute facial recognition. This suggests the need for more effective and efficient techniques that incorporate both labeled and unlabeled data for multi-attribute facial recognition.

3 Methodology

The proposed work is a semi-supervised multi-attribute facial recognition system, which predicts age, gender, ethnicity, and emotion based on facial images. We do not use a single coherent model; instead, we take a modular approach and train three specialized CNN models to predict attribute-specific features: (1) age and gender, (2) emotion, and (3) ethnicity. Figure 1 below **??**, shows that in our pipeline both labeled and unlabeled data have been employed in the training of a strong final model.

3.1 Datasets

To make sure that we have diversity and captured all four face features we use several datasets which are publicly available:

- **UTKFace:** A big data set which includes age, gender, and ethnicity labels.
- **FairFace:** In order to fight against demographic bias, we utilize the Fairface dataset [7] that comprises 108501 images evenly distributed across seven race groups and labeled by age and gender.
- **AffectNet:** It is an emotion recognition model that can be trained on a broad range of facial expressions in the wild and is used to teach computers how to recognize emotions[12].

Big data are processed by semi-automatic and automated methods of creating datasets. The semi-automatic TAAB as well as the simple annotation process on the facial landmark a were proposed by Sagonas et al. [14], and simple annotations do not also have to be ters markups and reduce the human labor. Tian et al. [15] and Zhang et al. [18] both proved the concept of interactive annotation by incorporating the unsupervised clustering with less user intervention to improve the labeling efficiency.Besides, automated pipelines [2] and synthetic data generation methods, including GAN-based approaches [4] and SynFace [6], have been shown to expand training datasets, improve model generalization, and mitigate data sparsity.

3.2 Preprocessing

The process of preprocessing all the facial images to stan- prepare the data to the neural network:

- **Face Detection and Alignment:** Faces are recognized and matched through normalized facial land- marks [14].
- **Normalization:** Pixels values are made normalized to the interval [0,1] to be able to feed the network.
- **Resizing:** Images are resized to a fixed input size (e.g., 48×48 or 64×64 depending on the model input requirements).
- **Data Augmentation:**Data augmentation methods, including horizontal flipping, rotation, brightness manipulation, and random cropping are used to make data more diverse and less overfitting.

3.3 Model Architecture and Training Strategy

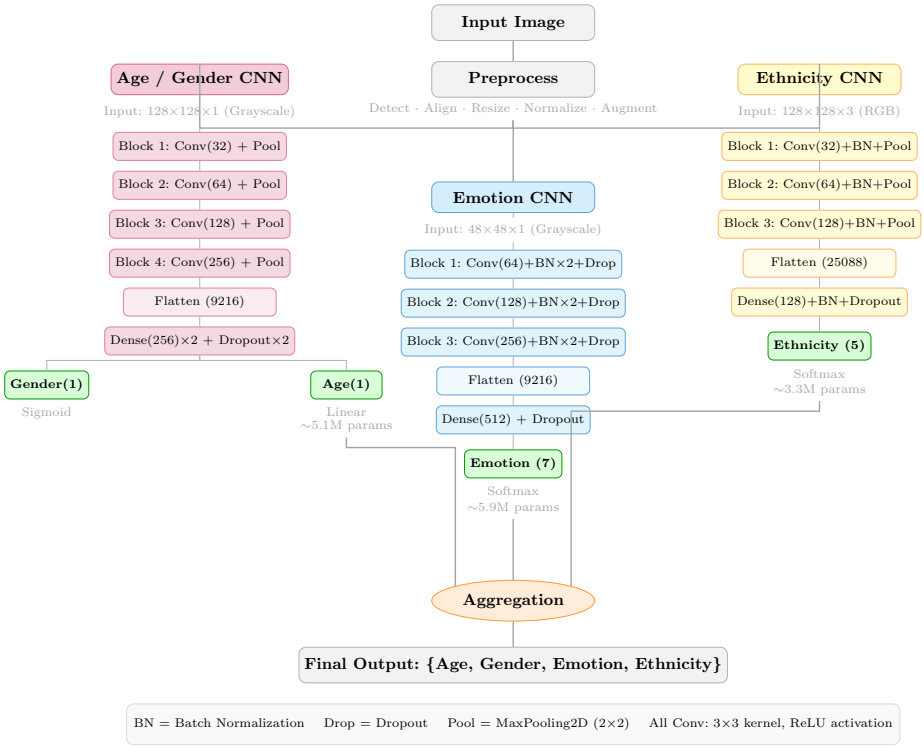


Fig. 1. Ensemble pipeline and internal architecture of the three custom CNN expert models. Each model is trained independently on its respective dataset before ensemble aggregation.

Our system's strength lies in a two-part strategy: (1) a semi-supervised pipeline to generate a massive, high-quality training dataset, and (2) an ensemble of custom-designed CNNs for final deployment.

Custom CNN Architecture Design Each of these models is built on a lightweight convolutional neural network (CNN) architecture which consists of multiple convolutional layers with 3 3 kernels, followed by Batch Normalization, ReLU activation, and Max Pooling layers. Extracted features are then fed through fully connected layers, and a Softmax layer is used to classify.

A single model to predict multi-attributes can be afflicted with conflicting representations of features in different tasks. Thus, to enhance the attribute-specific learning and the overall system performance, we take a modular multi-model approach.

Stage 1: Supervised Training of Custom Expert Models When training three special expert models we come first. Importantly, these models are first trained on the small labeled datasets (D_L) so as to form a basis of knowledge on their respective attributes. The stage 1 of the training is then succeeded by a more long-term stage that is the semi-supervised refiner process(Stage 2).

1. **Custom Age/Gender Model:** the model was Trained on the labeled portion of FairFace.
2. **Customized Emotion Model:** learned on the labeled portion of AffectNet.
3. **Custom Ethnicity Model:** custom ethnicity model was Trained on the labeled portion of FairFace.

Stage 2: Semi-Supervised Learning using FixMatch At this stage, unlabeled data is used with the FixMatch algorithm. Each unlabeled image is then first subjected to a weak augmentation to produce a prediction. When the prediction confidence surpasses a predefined threshold it is considered a pseudo-label. The model is then trained to make consistent forecasts when powerful augmentations of the same image are made. The process enhances the quality of pseudo-labels and enables effective use of unlabeled data.

Stage 3: Multi-Model Inference and Output Integration The last architecture implemented, as represented in Figure 1, is a combination of the three custom CNN models trained using the large, refined dataset ($D_L \cup D_{PL}$). This parallel architecture is better to use because of its modularity and strength.

1. **Parallel Inference:** An input image is preprocessed and fed simultaneously into the three independent Custom CNN models.
2. **Prediction:**Every model produces its high-confidence predictions of its particular work.
3. **Aggregation:** The individual forecasts are summed up into one and overall output vector:: {Image ID, Age, Gender, Emotion, Ethnicity}.

3.4 Training Procedure

- **Dataset Splitting:** The datasets are divided into training, validation and test sets, maintaining class balance between splits.
- **Optimization:** We make use of Adam optimizer and a learning rate schedule to enhance convergence.
- **Evaluation Metrics:** Accuracy, precision,recall, and F1-score is calculated To assess the diagnostic performance of system on test finate state sequence. computed as to each attribute.

3.5 Ethical Considerations

Under the consideration of sensitive attributes, ethnicity and age, the model is tested for demographic bias and fairness. Unfair training: balanced datasets such as FairFace [7] and Augmentation [4] decreases bogus predictions and makes sure of even more fair results for all groups of the population.

4 Research Objectives

The major objective of the work is to create scalable semi-supervised multi-attribute facial labeling system that can predict not only age, gender, emotion, or ethnicity but also all of them at the same time. The proposed model will seek to:

- Design a multiple output CNN architecture for different facial attribute prediction.
- Reduce manual labelling effort by refining pseudo labeles using FixMatch approach.
- resolve demographic bias using a balanced dataset such as FairFace [7].
- Enable real-time deployment for practical applications.

Our hypothesis is that expert models with semi-supervised refinement are more effective in enhancing the scalability of the data set and yet provide competitive classification performance.

5 Evaluation and Results

5.1 Performance Analysis

Each attribute was tested on the proposed multi-attribute system on held-out test sets. The most important measure of evaluation was accuracy, and precision and recall were taken as validation measures.

Table 1. Classification Accuracy for Each Attribute

Attribute	**Accuracy**
Age & Gender	0.89
Ethnicity	0.63
Emotion	0.68

The maximum accuracy (89 percent) was obtained with age and gender model, which means that the features were properly extracted, and the classification remained the same. The recognition of emotions had 68 percent accuracy which is as to be expected of facial expression differences. The classification of ethnicity was 63 percent, indicating that demographic prediction tasks are not that easy.

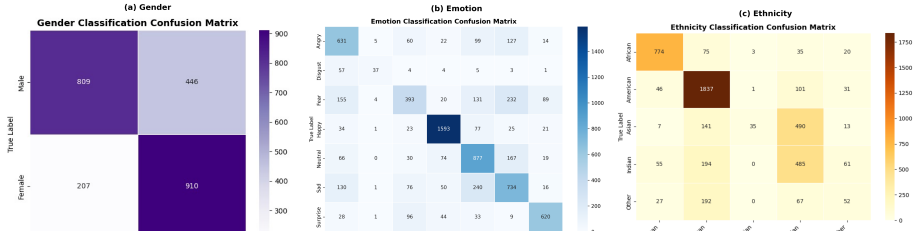
Table 2 draws a comparison between our proposed system and the existing methods of single attributes. Compared to these methods, which are based on individual facial properties, the proposed system uses multi-attributes prediction in parallel.

The comparison shows that although the single-attribute models prove to be more efficient in particular tasks, the given multi-attribute system shows

Table 2. Comparison with Existing Methods

Method	Dataset	Age	Gender	Emotion	Ethnicity
Levi & Hassner (2015) [9]	Adience	50.7%	86.8%	–	–
DEX / Rothe (2015) [13]	ChaLearn	MAE: 3.22	–	–	–
FairFace (2019) [7]	FairFace	–	94.89%	–	81.5%
AffectNet (2017) [12]	AffectNet	–	–	72%	–
UTKFace CNN	UTKFace	MAE: 3.8–5.7	91–98%	–	81–93%
Ours	Multiple	**89%**	**89%**	**68%**	**63%**

competitive outcomes and also allows promoting multiple attributes associated with the faces at the same time.



5.2 Observations

The confusion matrices of the three models is shown in Figure 5.1. The gender classifier is able to predict a large number of male and female samples correctly: 809 male and 910 female. For the recognition of emotion, the class Happy achieved the best accuracy (1,593) while the class Disgust achieved the lowest accuracy (37), as there are very similar expressions in this class. The model is able to classify American (1,837) and African (774) classes better while there is more confusion between Asian and Indian classes. In summary, most of the errors are made in similar and less represented visually challenging classes, which is an example of the difficulty of multi-attribute face recognition.

6 Conclusion and Future Work

This paper gives a complete mechanism of automatic labeling of faces that can determine ethnicity, emotion, age and gender simultaneously. The suggested system will address high performance and decrease demographic bias with the help of multi-output convolutional neural networks, transfer learning [13,9] ,and other datasets, such as UTKFace, FairFace [7], and AffectNet [12]. Semi-automatic [14] and interactive annotation systems [18,2] also improve the efficiency and quality of data labeling by means of reducing the number of people involved in the process, however, the reliability of the information remains unharmed. Additionally, synthetic datasets are suggested as a simple remedy to the training

data augmentation issue, and in situations where copyright or availability limits exist [4].

One is likely to assume that this assignment will lead to developing a strong multi- attribute facial recognition engine, the implications of unbalanced and synthetic data on fairness, and a demonstration of an effective annotation algorithm on large data. The whole methodology has stressed ethics to ensure that there is responsible use of AI models that handle sensitive attributes.

6.1 Future Work

A number of directions could further develop and extend the proposed system:

- **Real-Time Deployment:**To allow real time prediction on edge devices or mobile platforms optimize lightweight and different models.
- **Extended Attributes:** added facial attributes, e.g. facial hair, glasses, or pose variations, to improve the understanding of the model.
- **Dataset Generalization:** incorporate additional large-scale datasets and fairness-aware training techniques to further improve generalization across demographic groups [7,17,3,19].
- **Semi-Supervised Learning:** Explore semi-supervised and self-supervised learning strategies to handle unlabeled data, which also reduce the reliance on labeled datasets [2,4].
- **Bias Mitigation:** Build more advanced fairness-conscious training methods to reduce residual demographic or gender bias in predictions.
- **Ethical Auditing:** select methodical audits and user testing to evaluate social and ethical implications of facial recognition multisensory systems.

Focusing on these directions of the future, the proposed framework can transform into a fully-fledged and accountable system of facial analysis that can be used in the context of security, social media, virtual customer service, and research.

References

1. Albiero, V., Bowyer, K.W., Vangara, K., King, M.C.: Analysis of gender and race bias in face recognition: A case study on the passageface dataset. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1950–1959 (2021). https://doi.org/10.1109/WACV48630.2021.00199
2. Anvari, Z., Athitsos, V.: A pipeline for automated face dataset creation from unlabeled images. In: ACM International Conference Proceeding Series. pp. 227–235. Association for Computing Machinery (6 2019). https://doi.org/10.1145/3316782.3321522
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018). pp. 67–74 (2018). https://doi.org/10.1109/FG.2018.00020

4. Colbois, L., Pereira, T.D.F., Marcel, S.: On the use of automatically generated synthetic image datasets for benchmarking face recognition. In: 2021 IEEE International Joint Conference on Biometrics, IJCB 2021. Institute of Electrical and Electronics Engineers Inc. (8 2021). https://doi.org/10.1109/IJCB52358.2021.9484363
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698 (2019)
6. Gong, S., Wang, W., Liang, X., Deng, J., Zafeiriou, S.: Synface: Face recognition with synthetic data. arXiv preprint arXiv:2108.07025 (2021)
7. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age (8 2019), http://arxiv.org/abs/1908.04913
8. Kortylewski, A., Schneider, A., Gerig, T., Egger, B., Luan, V., Vetter, T.: Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
9. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2015)
10. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 212–220 (2017)
11. Merler, M., Lokhande, R.T., Srivastava, A., Smith, J.R., Feris, R.S.: Diversity in faces. arXiv preprint arXiv:1901.10436 (2019)
12. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
13. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops (December 2015)
14. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 896–903 (2013). https://doi.org/10.1109/CVPRW.2013.132
15. Tian, Y., Liu, W., Fang, R.X., Tang, W.X.: A face annotation framework with partial clustering and interactive labeling. Tech. rep.
16. Wang, H., Wang, Y., Zha, Z.J., Zhu, Y.M., Sun, W.W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
17. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 529–534 (2011)
18. Zhang, L., Tong, Y., Ji, Q.: Lncs 5303 - active image labeling and its application to facial action labeling. Tech. rep. (2008)
19. Zhu, J., Wu, S.F., Zhao, S.L., Liu, Z.T., Li, Y., Zhou, Z.H.: Celebv-hq: A large-scale video facial attributes dataset. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 4121–4130 (2022). https://doi.org/10.1145/3503161.3547942