Cultural Context-Aware Facial Expression Recognition Using Latent Space Clustering

Cynthia Obianuju Akanaga
akana22c@mtholyoke.edu

***Abstract*— Facial Expression Recognition (FER) is an exciting field that aims to come as close as possible to the ability of humans to decipher emotions using simply an image of a person's face. Despite advancements, these systems are often biased due to imbalanced datasets, raising concerns about their fairness. This paper investigates demographic biases in a notable FER dataset, focusing on latent space clustering patterns to reveal disparities in racial groups. Using the UTKFace dataset and ResNet50 model for feature extraction, I applied K-Means clustering to identify patterns in demographic representation. Results show significant underrepresentation of Black individuals, across clusters, underscoring the influence of dataset biases on FER performance. Next steps would be to investigate correlations between other labels (e.g., age, gender) and clustering patterns to uncover multidimensional biases. Also to implement the clustering on other datasets like FairFace and try integrating the underrepresented group into the each cluster to form a balanced set.**

***Index Terms*— Facial Expression Recognition; Latent Space; Feature Extraction**

I. Introduction

Facial expressions may be the universal language of emotion, as theorized by Paul Ekman, a renowned psychologist. Ekman identified seven basic human emotions — happiness, sadness, anger, fear, surprise, disgust, and contempt — that are universally expressed through facial expressions across cultures [1]. However, what happens when the algorithms designed to classify these expressions are unable to accurately read faces for certain demographics? Is it still universal?

Facial Expression Recognition (FER) is a subfield within Affective Computing, the area of artificial intelligence dedicated to understanding and interpreting human emotions to create more personified machines [2]. FER uses patterns in images of people's faces, such as a furrowed brow or a slight pout, to identify emotions like happiness, sadness, anger, or surprise. By analyzing facial expressions to infer emotions, FER offers valuable insights that can be used in diverse fields, from healthcare to autonomous vehicles. However, FER systems are only as good as the datasets they are trained on. The datasets utilized for FER algorithms can be categorized into two types: lab-based and in-the-wild datasets [3].

A. Types of FER Datasets

FER datasets are categorized based on the environment in which the data is collected [3]:

1) Lab-based Datasets: Obtained in controlled environments with consistent lighting, background, and camera angles. Two notable examples are CK+ (Extended Cohn-Kanade dataset and JAFFE dataset). These are unrealistic for real-world applications, as people in everyday scenarios rarely sit up straight, face the camera directly, or pose expressions in ideal lighting conditions.

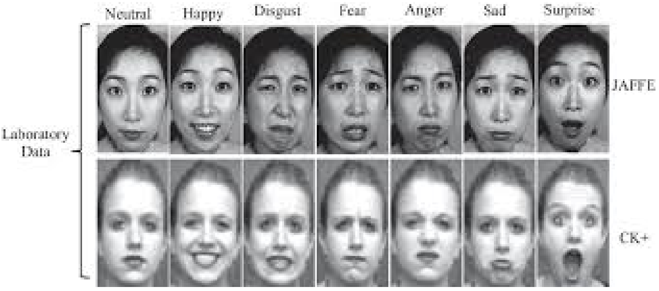

Fig. 1. Example of a Lab-based dataset from JAFFE and CK+ [4], [5].

2) In-the-Wild Datasets: Obtained in natural, uncontrolled environments with diverse lighting, poses, and occlusions (e.g., glasses, masks). An example is the UTKFace dataset, which I used in my running code. These are messier images with challenges like dim lighting, obstructions to the face, and variations in image quality. But it is more realistic and representative of real-world conditions, making it suitable for real-world applications.

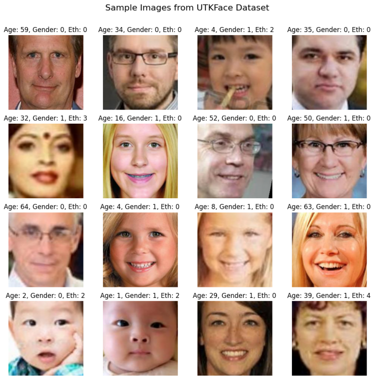
Sample Images from UTKFace Dataset
Age: 59, Gender: 0, Eth: 0
Age: 34, Gender: 0, Eth: 0
Age: 4, Gender: 1, Eth: 2
Age: 35, Gender: 0, Eth: 0
Age: 32, Gender: 1, Eth: 3
Age: 16, Gender: 1, Eth: 0
Age: 52, Gender: 0, Eth: 0
Age: 50, Gender: 1, Eth: 0
Age: 64, Gender: 0, Eth: 0
Age: 4, Gender: 1, Eth: 0
Age: 8, Gender: 1, Eth: 0
Age: 63, Gender: 1, Eth: 0
Age: 2, Gender: 0, Eth: 2
Age: 1, Gender: 1, Eth: 2
Age: 29, Gender: 1, Eth: 0
Age: 39, Gender: 1, Eth: 4

Fig. 2. Example of an in-the-wild dataset from UTKFace [6].

Facial Expression Recognition (FER) systems typically follow three steps: data preprocessing, feature extraction, and classification. The specific methods used for these steps vary depending on whether traditional techniques or deep learning approaches are applied [3].

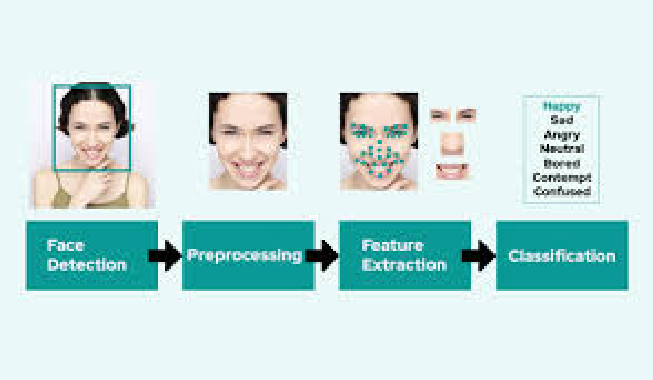

Fig. 3. General FER process [7]

B. Approaches in Facial Expression Recognition

1) Traditional Methods: Traditional methods rely on simple classifiers. The data preprocessing often involves aligning facial images to ensure uniformity, such as detecting landmarks like the eyes, nose, and mouth. Feature extraction uses a descriptor like Local Binary Patterns (LBP) to encode facial expressions into vectors. Finally, classifiers such as Support Vector Machines (SVM) are used to categorize emotions based on these features. [3]

These methods are quick and work well for small datasets, but struggle to generalize to in-the-wild datasets due to their inability to capture complex and diverse patterns in facial expressions.

2) Deep Learning Methods: Deep learning methods automate much of the FER pipeline. The data reprocessing is simplified to basic operations like resizing and normalizing images to prepare them for input into a neural network. Feature extraction is performed automatically by convolutional layers in a deep neural network, such as ResNet50 (I used this for my code) or VGG16, which learn hierarchical representations of facial features directly from data. The extracted features are then passed through fully connected layers to classify emotions. [3]

Deep learning methods outperform traditional techniques in accuracy and generalization, especially on large and diverse datasets. However, with more innovation comes new problems. Since much of the process is automatic, it runs into the problem of making sure the model is learning the right features for expression classification. They are also quite expensive to run and need large amounts of labeled data for effective training. Most of the recent studies in FER utilized teh deep learning method.

C. Key Concepts and Variables for this study

This study employs a clustering-based approach to analyze biases in FER systems. To support this, several key concepts and variables are used:

1) Latent Space: Latent space refers to a high-dimensional feature space where facial images are represented as vectors. These features are extracted from facial images using a deep learning model, such as ResNet50, and capture essential patterns in the data.

2) Clustering with K-Means: Clustering is a method of grouping data points based on their similarity in latent space. This study employs the K-Means algorithm to analyze the latent features of faces. K-Means assigns each image to a cluster, enabling the analysis of how different demographic groups are represented within these clusters.

3) Bias in FER: Bias refers to systematic discrimination in the performance of FER systems across demographic groups. Bias often arises when training datasets are imbalanced or underrepresent certain groups.

II. Existing Methods

In the field of Facial Expression Recognition (FER), both traditional and deep learning methods have faced challenges such as sensitivity to differences in lighting, non-frontal poses, occlusions, and the need for extensive labeled data. Recent research has focused on addressing these limitations to enhance the accuracy and robustness of FER systems.

1) Enhance feature representation: Liao et al. introduced RCL-Net, which combines Local Binary Patterns (LBP) with a residual attention network. This architecture emphasizes local facial details and integrates channel and spatial attention mechanisms to enhance feature representation, improving recognition accuracy in both controlled and wild environments. [3]

2) GAN (Generative Adversarial Networks): Zhang et al. proposed an end-to-end deep learning model based on Generative Adversarial Networks (GANs). Unlike traditional methods that rely on face frontalization or pose-specific classifiers, this model jointly learns to synthesize facial images and perform pose-invariant FER. This is particularly relevant to bias in FER, as it reduces the risk of performance disparities across demographic groups by enriching the training set with more diverse examples. [8]

3) Transformer-based models: Li et al. introduce **FER-former**, a Transformer-based approach designed to advance FER in uncontrolled environments. The method leverages a hybrid stem to combine the strengths of CNNs and Transformers, enabling simultaneous use of image-based features and text-oriented tokens for classification. To address annotation ambiguity, FER-former supervises the similarity between image and text features, aligning image semantics with text-space representations. [9]

III. Open Challenges

A. Current Challenges

Demographic bias remains a significant challenge in Facial Expression Recognition (FER), as many systems fail to account for cultural and individual variations in emotional expressions. While foundational works, such as those by Ekman, have posited the universality of emotions, recent studies highlight the limitations of this assumption when applied to diverse populations [1]. Addressing these concerns, [10] undertook a systematic investigation of bias and fairness in FER systems. They compare baseline, attribute-aware, and disentangled approaches using well-known datasets (RAF-DB and CelebA) and find that the disentangled approach, particularly when combined with data augmentation, is the most effective for mitigating demographic bias. Their findings underscore the importance of designing models that prioritize fairness alongside accuracy, especially in the context of uneven attribute distributions or imbalanced subgroup data.

Despite progress in bias mitigation, open challenges remain. Most facial expression datasets, such as RAF-DB, are skewed toward certain demographic groups, predominantly Caucasian individuals within a narrow age range. Additionally, as noted in [10], larger datasets like CelebA may not benefit significantly from existing augmentation techniques, pointing to the need for alternative approaches to mitigate bias in well-represented data. Many existing approaches, while effective in controlled settings, fall short for in-the-wild datasets.

B. Research Ideas

Given the challenges in mitigating bias in Facial Expression Recognition (FER), my research focuses on leveraging latent space analysis for unsupervised bias detection using the UTKFace dataset. My research aims to identify demographic biases in the model's latent feature space. By analyzing cluster distributions and measuring overlaps, the goal is to reveal subtle biases that may not be captured through conventional fairness metrics.

IV. Concept to Code

A. Dataset

The dataset used for this study is the **UTKFace dataset**, sourced from a publicly available repository. It contains facial images annotated with the following labels:

- **Age:** Numerical values representing the age of the individual.
- **Gender:** Binary values (`0` for male, `1` for female).
- **Ethnicity:** Categorical values corresponding to racial categories (`0` = White, `1` = Black, `2` = Asian, `3` = Indian, `4` = Other, including Hispanic, Latino, and Middle Eastern).

Key Characteristics:

- **Diversity:** The dataset spans a wide range of ages (0–116 years), genders, and racial categories. However, the representation of racial groups is imbalanced:
  - White individuals dominate the dataset (**54.2%**).
  - Black individuals are severely underrepresented (**2.6%**).
- **Preprocessing:** Images are provided in cropped and aligned formats, ensuring that facial features are centralized for analysis.
- **Size:** The dataset contains over 20,000 images. For computational feasibility, a subset of 500 images was used in this project.

The FairFace dataset would have been more ideal, but I decided to use UKFace instead because I struggled to use a dataset that had the labels in a different folder. In the UTKFace dataset, the images are named according to their race, gender and age so I don't need to download anything else, just the images so it was easier and faster to load into Colab.

B. Running the Code

The code can be run by going to the following GitHub folder: `https://github.com/MHC-FA24-CS341CV/beyond-the-pixels-emerging-computer-vision-research-topics-fa24/blob/main/code/06-facial-expression-recognition/06_facial_expression_recognition.ipynb` and downloading the file. Then upload it into google colab and select ”Run All” from the Runtime dropdown. Then you can scroll through and view the images and conclusions from the kmeans clustering

C. Results

Using the K-Means algorithm with $n = 4$ clusters, the facial features extracted from the UTKFace dataset were grouped into distinct clusters. Each cluster represents a grouping of similar facial features based on the latent feature space generated by the ResNet50 model. The clusters were analyzed to uncover patterns in racial distribution and potential biases. Key observations:

TABLE I
Ethnicity Distribution Across Clusters

Cluster	White	Black	Asian	Indian	Other
0	58.8%	3.0%	14.1%	13.1%	11.1%
1	51.0%	3.0%	20.0%	14.0%	12.0%
2	36.8%	0.0%	36.8%	21.0%	5.3%
3	52.7%	2.2%	15.9%	18.6%	10.4%

Cluster 0: This was the largest cluster, containing 58.8% White individuals, 14.1% Asian individuals, 13.1% Indian individuals, and smaller proportions of Other (11.1%) and Black (3.0%) individuals. The dominance of White individuals in this cluster reflects the overall dataset bias toward this group.

Cluster 1: This cluster had a more balanced composition compared to others, with 51.0% White individuals, 20.0% Asian individuals, 14.0% Indian individuals, and 12.0% Other individuals. Black individuals were again underrepresented at 3.0%.

Cluster 2: This was the smallest cluster, containing 36.8% White individuals, 36.8% Asian individuals, and 21.0% Indian individuals. Black individuals were absent from this cluster, and Other individuals made up a small proportion (5.3%). The low representation of White individuals in this cluster suggests it captured unique or outlier features not common in the dataset.

Cluster 3: This cluster had 52.7% White individuals, 18.6% Indian individuals, 15.9% Asian individuals, and 10.4% Other individuals. Black individuals accounted for only 2.2%, continuing the trend of underrepresentation.

Cluster 4: Similar to other clusters, White individuals dominated at over 50%, with smaller proportions of Asian, Indian, and Other individuals. Black representation was consistently low.

D. Conclusions

1. Dataset Bias: The clustering results highlight significant imbalances in the dataset. White individuals dominate all clusters, reflecting their overrepresentation in the UTK-Face dataset (54.2%). Black individuals, on the other hand, are severely underrepresented across all clusters, making up only 2.6% of the dataset.

2. Model Bias: The clustering patterns reveal that the latent feature space learned by ResNet50 is influenced by dataset biases. The dominance of White individuals in most clusters suggests the model prioritizes features common to this group, at the expense of underrepresented groups like Black individuals.

V. Future Work

Dataset Balancing: Augment the dataset with additional images of Black individuals and other underrepresented groups to ensure equitable representation.

References

[1] P. Ekman, *Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions*. Los Altos, CA: Malor Books, 1975.

[2] D. Team. (2022) What is affective computing? Accessed: 2024-11-23. [Online]. Available: https://www.datacamp.com/blog/what-is-affective-computing

[3] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial expression recognition methods in the wild based on fusion feature of attention mechanism and lbp," *Sensors*, vol. 23, no. 9, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/9/4204

[4] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "The japanese female facial expression (jaffe) database," 1998, accessed: 2024-11-23. [Online]. Available: https://zenodo.org/record/3451524

[5] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, 2000, accessed: 2024-11-23.

[6] Z. Zhang, "Utkface: A large-scale dataset of faces with age, gender, and ethnicity labels," 2017, accessed: 2024-11-23. [Online]. Available: https://www.kaggle.com/datasets/jangedoo/utkface-new

[7] A. Biometrics, "How does facial emotion recognition express your feelings?" 2022, accessed: 2024-11-23. [Online]. Available: https://www.aratek.co/news/how-does-facial-emotion-recognition-express-your-feelings

[8] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3359–3368.

[9] Y. Li, M. Wang, M. Gong, Y. Lu, and L. Liu, "Fer-former: Multi-modal transformer for facial expression recognition," 2023. [Online]. Available: https://arxiv.org/abs/2303.12997

[10] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 506–523.