CelebA Dataset Detailed Report

Overview

Purpose: The CelebFaces Attributes (CelebA) dataset is designed to support machine learning tasks in facial recognition, attribute prediction, face detection, and generative adversarial network (GAN) training. It provides diverse and extensive annotations for multiple facial attributes.

Creators: The dataset was created by Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang as part of the Multimedia Laboratory (MMLab) at The Chinese University of Hong Kong. It was first introduced in their work, "Deep Learning Face Attributes in the Wild."

Applications: CelebA is widely utilized in:

- Facial attribute recognition (e.g., gender, emotion, age).
- Benchmarking face detection algorithms.
- Training and evaluation of generative models like GANs.
- Style transfer tasks in computer vision.
- Research into bias and fairness in AI systems.

Dataset Composition

General Statistics

- **Number of Images:** 202,599 high-resolution face images.
- **Number of Unique Identities:** 10,177 celebrity identities, covering a diverse range of ethnicities, ages, and gender representations.
- **Image Dimensions:** All images are aligned and cropped to 178x218 pixels.

Facial Attributes

Each image is annotated with 18 binary attributes. These include:

- **Appearance-based attributes:** Bald, Bangs, Black Hair, Blond Hair, Brown Hair, Gray Hair, Receding Hairline, Straight Hair, Wavy Hair.
- **Accessories:** Eyeglasses, Earrings, Hat, Necklace, Necktie.

- **Expressions and Actions:** Smiling, Blurry, Mouth Slightly Open.
- **Facial Features:** Arched Eyebrows, Bushy Eyebrows, Narrow Eyes, Oval Face, High Cheekbones, Chubby.
- **Presence of Facial Hair:** Beard, Goatee, Mustache, Sideburns.
- **Other Attributes:** Attractive, Young, Wearing Lipstick, Male.

Landmark Annotations

The dataset includes five key landmark points for each face:

- Left eye
- Right eye
- Nose tip
- Left mouth corner
- Right mouth corner

Partitioning

The dataset is divided into three subsets for standardized training, validation, and testing:

- **Training set:** 162,770 images.
- **Validation set:** 19,867 images.
- **Testing set:** 19,962 images.

Use Cases and Benchmarks

1. Facial Recognition: CelebA is extensively used to train and benchmark facial recognition models due to its large volume and diversity.

2. Attribute Prediction: Researchers use CelebA for multi-label classification tasks, predicting attributes such as gender, hairstyle, and expression.

3. Face Detection: CelebA provides a standardized benchmark for evaluating face detection algorithms.

4. Generative Models: CelebA serves as a primary dataset for training GANs to generate realistic human faces or modify attributes in existing images (e.g., adding glasses or changing hair color).

5. Ethical AI Research: The dataset is used in studying bias in machine learning models, particularly around representation across gender and ethnic groups.

Advantages and Limitations

Advantages

- Large scale with diverse identities and attributes.
- Provides both binary attributes and precise facial landmarks.
- Widely recognized as a benchmark dataset in computer vision.

Limitations

- Celebrities may not represent real-world population diversity.
- Binary attribute labels can oversimplify certain characteristics.
- Potential for ethical concerns regarding consent and fairness in AI models.

References and Resources

- Official Website: `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`
- Paper: `https://arxiv.org/abs/1411.7766`
- Dataset Access: `https://www.kaggle.com/jessicali9530/celeba-dataset`