Differential Privacy in Facial Recognition

Introduction

In the era of ubiquitous facial recognition technology, the preservation of individual privacy has become a paramount concern. Particularly, facial recognition presents risks of identity disclosure. Biometric information is immutable and unique to an individual, and thus can cause irreversible harm if this data were released or leaked to a third party. Governments and private industry continue to embrace facial recognition for its relative security and convenience benefits compared to the traditional authentication approaches of passcodes and passwords. Facial recognition can be up to 20 times more secure than fingerprint-based identification, at least in relation to smartphone usage (Galov 2023). Additionally, some recognition algorithms boast accuracy rates of 99%+ (OLOID 2023), further pushing adoption.

However, the court of public opinion has mixed feelings on the use of facial recognition (Ritchie et al., 2021). Regardless, use of facial recognition technology may one day be practically unavoidable around the world, and is already adopted at different levels in different countries. For example, since 2017, customers can use “smile to pay” technology at KFC restaurants in China (Russell 2017), a concept likely unthinkable to countries earlier in facial recognition adoption lifecycle.

This project aims to reproduce the results of the paper titled "Privacy-Preserving Face Recognition with Learnable Privacy Budgets in Frequency Domain" (Ji et al. 2022). Our goal is to perturb the human-readable aspects of a person’s facial image and produce a machine-readable version of the image for recognition, limiting information leakage while still preserving utility. The goal being enhancing privacy, while maintaining the utility and security that facial recognition offers.

Survey of Related Work

Privacy in Facial Recognition

Homomorphic encryption has been used in privacy-protecting research to process and encrypt image data. This encryption method relies on a semi-honest threat model that provides the capability to encrypt data in ways that preserve privacy, but has a high computational cost which limits how well it scales. PEEP (privacy-preserving face recognition protocol) aimed to address the issue with high computational complexity of Homomorphic encryption, but it reduced the overall recognition quality (Chamikara et al. 2020). We will attempt to address the computational complexity and recognition quality concerns with our approach.

The most relevant work related to ours is the work of Ji et al. (2022), which we aimed to reproduce as part of this study. Ji et al. claimed to introduce a novel approach to privacy preserving facial recognition by developing learnable privacy budgets during training, and demonstrated that utility was not sacrificed in the process. This was shown by comparing their model accuracy to other similar work, where they claimed to have better accuracy and privacy.

Other models have used Gaussian noise distribution instead of Laplacian, which is the method that we used in our differential privacy approach. While effective, Gaussian noise often requires a delicate balance as it can easily either over-smooth or under-protect the image, potentially leading to either loss of recognition capability or insufficient privacy protection (TutorMaster 2023).

Some systems opt for simpler methods like pixelation or blurring (Fan 2018). While user-friendly and straightforward, these techniques are less secure against advanced de-anonymization algorithms and can significantly reduce the utility of images for recognition purposes.

The Generative Adversarial Network (GAN) Based Models are more complex and involve training a neural network to generate noise or alter images in a way that preserves privacy (Aggarwal et al. 2021). While highly effective and customizable, GAN-based models require extensive training data and computational resources, and can sometimes produce artifacts that reduce image utility.

Image Processing and Compression

Discrete Cosine Transform (DCT), the method we used to remove the low-frequency information from our images, has also been used as an effective way for image storage via compression for JPEG images. Cabeen and Gent demonstrate how images can be compressed for storage using DCT, and then subsequently decompressed for processing with limited impact to image quality and recognition. This paper exhibits the merits of being able to translate the image from one domain to another.

Datasets

For our work, we utilized two publicly available datasets. In both instances, we used the data as it was provided, and did not make additional changes that would require elevated permissions. The first dataset, called VGG-Face2, contains 3.31 million images of 9131 subjects, with an average of 362.6 images for each subject. (Qiong 2018). This dataset was used for Exploratory Data Analysis (EDA) and for initial privacy model development, but was not used in our machine learning model.

For our machine learning model, we utilized images from the Ethnicity Aware Training Dataset (Wang, Deng 2020), which attempts to have balanced data between different racial groups. In this data, there are four distinct categories, with an equal representation among them: Caucasian, African, Asian, and Indian. We recognize that this is a limited representation of the world's population and racial identities, and does not account for the non-binary nature of race. Additionally, there are concerns with the verbiage used, and the potential methods used for classification. However, with known biases found in facial recognition (Buolamwini, Gebru 2018), we chose this dataset so that we could include different skin complexions in our models, even if the representation was limited and generalized to only four groups. Having a more diverse group of input images allowed us to explore the ability of DCT to obscure images, regardless of the many variations of facial features. During training, fictional names were used for each individual's label to forgo any risks associated with labeling on race.

Methods

Privacy Preservation

Overview

The chosen model for noise perturbation, which employs Discrete Cosine Transform (DCT) and the Laplace noise distribution (differential privacy), was selected due to its unique balance of privacy preservation and image usability. This model stands out for several reasons:

- **Optimal Balance Between Privacy and Recognition:** The key strength of this model is its ability to maintain the utility of images for facial recognition while adequately masking personal identity (Ji et al. 2022). By focusing noise addition on high-frequency components, it ensures that the core facial structure needed for recognition is preserved, while personal identifiers understandable by humans are obscured.
- **Adaptive Noise Application:** This model's adaptive algorithm, which applies differential noise levels based on the sensitivity of facial regions, allows for more nuanced privacy protection, ensuring that more identifiable areas receive a higher level of noise perturbation.
- **Robustness Against Re-Identification Attacks:** By employing Laplacian noise and DCT based direct component (DC) channel pruning, this model provides a higher degree of protection against sophisticated re-identification techniques. The mathematical properties of Laplacian noise make it more challenging for attackers to reverse-engineer or filter out the noise to retrieve the original image. This, in addition to the removal of low frequency DC channel, significantly reduces the risk of re-identification.

Implementation

Image Preprocessing

The first step in our implementation was to scale the input range from [-1, 1] to [0, 255] to map the pixels to a standard RGB image range. We then converted the images from the RGB color space to YCbCr by subtracting 128. This new input range [-128, 127], is required in order to conduct our next step.

Discrete Cosine Transform (DCT)

DCT is used to convert the images from the image domain to the frequency domain. This is done to identify and process different parts of the image based on frequency. The human eye relies on low-frequency information for image recognition, while neural networks can use both low and high-frequency information for recognition (Ji et al. 2022). Thus, performing DCT is our first step in making the image unrecognizable to the human eye, by removing the low-frequency information from an image. The two-dimensional DCT of an image is calculated using the following formula:

$$DCT(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos\left[\frac{(2x+1)u\pi}{2M}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right]$$

where $\alpha(u)$ and $\alpha(v)$ are scaling factors, and $M \times N$ is the size of the image.

Differential Privacy

Traditional differential privacy uses adjacency in databases and query sensitivity. The paper relies on a slightly different mechanism to leverage differential privacy for images. It considers the BDCT version of

the image to be the secret and relies on the distance between secrets instead of the adjacency of databases. The distance metric would ensure that similar looking secrets were indistinguishable, while very different secrets still remain distinguishable.

The min and the max values for each point in the image were recorded as sensitivities for the image. The distance was then calculated using the following formula:

$$d_{i,j,k}(x_1, x_2) = \frac{|x_1 - x_2|}{r_{max}^{i,j,k} - r_{min}^{i,j,k}} \qquad \forall x_1, x_2 \in R_{i,j,k}$$

Noise Addition

Inspired by Ji et al., we applied Laplacian noise to high-frequency components outputted from the DCT image. The noise, drawn from the Laplace distribution with a mean (μ) and a tunable scale parameter (b), is calculated using:

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

This step ensures the obfuscation of finer details while retaining the overall image structure. Since we already removed the low-frequency components from the image, limiting human-readable components, we decided to set μ to 0, and b to the size of the input. In our context, this means that our noise is centered on the DCT image, and the scale of the distribution is learned throughout the training process.

Considerations when applying noise

- **Scale Parameter Tuning (b)**: Balancing the scale parameter to maintain a trade-off between privacy and image utility for recognition.p
- **Balancing Recognition and Privacy**: The objective is to ensure sufficient noise for privacy without overly compromising facial recognition utility.

CNN Model

To evaluate the utility of images after implementing DCT and differential privacy, we developed a convolutional neural network (CNN) model. This model served two purposes:

1. **Baseline Recognition**: Initially, the CNN model performed facial recognition on original images without privacy enhancement to establish a baseline for measuring utility.
2. **Recognition Post Privacy Enhancement**: The same CNN was then trained on original images and tested on perturbed images to compare recognition performance.

Our model split our data to use 80% for training and 20% for testing. The training utilized the Adam optimization algorithm with a learning rate of 0.001 over 10 epochs. Due to resource constraints, the training involved only 733 images from 12 individuals, representing three people from each racial group, as identified in the Ethnicity Aware Training Dataset.

Results

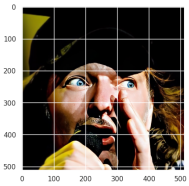
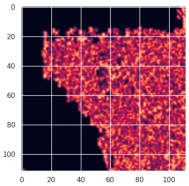
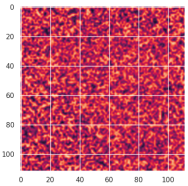
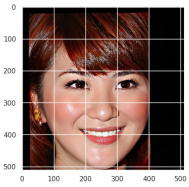
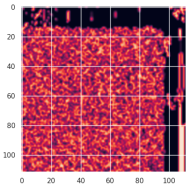
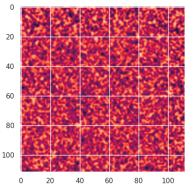
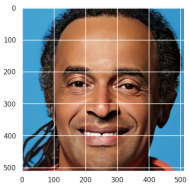
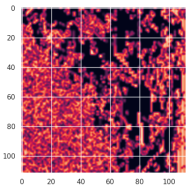
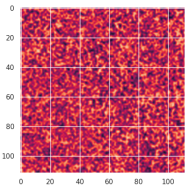
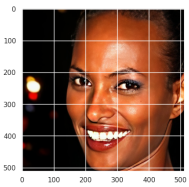
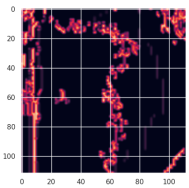
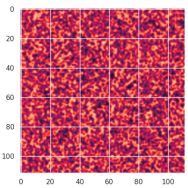
Privacy Preservation

Visual Evaluation

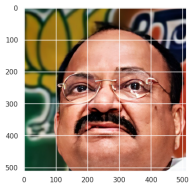
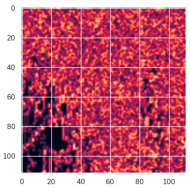
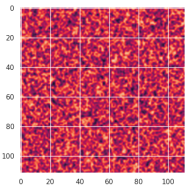
In Table 1 below, we observe the following:

- **Original vs. Post-DCT Images**: The transition from the original to the post-DCT image demonstrates a significant reduction in recognizable features, with the DCT effectively blurring or distorting distinct facial characteristics.
- **Post-DCT vs. Noise-Added Images**: The addition of Laplacian noise further obscures identifiable features. The noise application, targeted primarily at high-frequency components, makes it challenging to discern individual facial details, thereby enhancing privacy.

Table 1: Privacy Preservation: Examples of Images

Original Image	Post-DCT	Post-DCT with Noise
		
		
		
		

Original Image **Post-DCT** **Post-DCT with Noise**

Technical Analysis of Noise Efficacy

- **Frequency Component Alterations**: Post-DCT, the high-frequency components where the noise was added showed substantial alterations. These modifications played a key role in preserving privacy while maintaining the structural integrity required for facial recognition. As part of the DCT module, low frequency DCs were removed from the image. This makes it very difficult for these images to be fully reconstructed.
- **Noise Distribution Characteristics**: Analyzing the noise distribution, we observed that the Laplacian noise was effective in creating a significant barrier against potential reconstruction of the original image. The scale parameter 'b' appears to have sufficiently introduced enough noise to make the image unrecognizable.

While the DCT stage effectively removed human-level readability, as seen in Table 1, the noise activation is essential for privacy preservation as it eliminates the possibility of reconstructing the image to its original form.

Utility Assessment

Model Performance Metrics

- **Accuracy Reduction**: The accuracy of recognizing individuals from the privacy-enhanced images was 13.51%, a considerable reduction from the baseline model's 72.58% accuracy.
- **Recognition Capability**: The significant reduction in accuracy indicates that, while the model maintains some level of recognition capability, the added noise impacts its ability to accurately identify individuals compared to baseline.

Comparison and Interpretation

- **Baseline vs. Privacy-Enhanced**: The stark contrast in accuracy between the baseline and privacy-enhanced images underscores the trade-off between privacy preservation and utility. The results suggest that while our approach is successful in enhancing privacy, it does so at the cost of reduced recognition accuracy.
- **Random Guessing Comparison**: The fact that model performance was only marginally better than random guessing (8.33%) for a 12-option scenario indicates the substantial impact of our privacy-preserving techniques on the utility of the system. This does, however, indicate that we were able to preserve some level of utility, since recognition was better than random guessing.

Future Directions for Utility Improvement

- **Dataset Expansion**: Future research will involve expanding the dataset beyond the current 733 images to provide a more extensive evaluation of the model's utility.
- **Algorithm Optimization**: We aim to refine the noise addition algorithm, possibly through adaptive noise application or more sophisticated DCT coefficient analysis, to achieve a better balance between privacy and recognition accuracy.
- **Deep Fakes:** In future research, we hope to test our model against deep fake images in order to assess the ability of facial recognition to discern between real images and AI generated images.

Discussion

In our research, we were able to introduce DCT and implement differential privacy with learnable Laplace distribution scaling parameters to perturb image data and aid in privacy-preserving facial recognition technology, as done by Ji et al. Unfortunately, we were unable to replicate the accuracy found in their report, potentially due to our limited training data. We were, however, able to achieve similar results for preserving privacy, especially for the goal of making images harder for humans to recognize.

With this knowledge that privacy can be preserved, individuals who are not comfortable using facial recognition may instead adopt this technology for personal use. Without the ability for attackers to gain access to accounts or other personal effects via passcodes or passwords, often not secure (Binckes 2023), increased use of facial recognition for authentication could result in a reduction in cyberattacks and identity theft across industries.

This approach to facial recognition could also be used to aid in the reduction of bias when facial recognition is used for policing. As police precincts across the United States turn to facial recognition to offset the reduction of police officers (Khalil 2023), there have been many instances of wrongful arrests and imprisonment, mainly impacting marginalized communities (Hill 2020). While biases within facial recognition algorithms themselves continue to raise concerns and be an ongoing challenge, the methods shared in this research could potentially limit the amount of human bias that may accompany these systems. In other words, if a police officer does not see a human-recognizable face until after accuracy from a model with perturbed images meets a certain threshold, we could rely more on the software to recognize individuals, while preserving the privacy of individuals who are not deemed a match.

With the continued adoption of facial recognition, it may soon be inevitable that this form of encryption is required to access our phones, bank accounts, places of employment, etc. It is our hope that firms consider the privacy implications of this technology along with the added security and utility, and weigh these three factors equally moving forward.

References

AI TutorMaster. (2023, January 18). *Thought leadership from the most innovative tech companies, all in one place. What is Gaussian Noise in Deep Learning? How and Why it is used?* https://plainenglish.io/blog/what-is-gaussian-noise-in-deep-learning-how-and-why-it-is-used

Alankrita Aggarwal, Mamta Mittal, & Gopi Battineni. (2021). *Generative adversarial network: An overview of theory and applications*. https://www.sciencedirect.com/science/article/pii/S2667096820300045

Elise Devaux. (2022). What is Differential Privacy: Definition, mechanisms, and examples. *Statice.Ai*. https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples#:~:text=Definition%20of%20differential%20privacy,any%20individual%20in%20the%20dataset

Jahd Khalil. (2023, August 16). Real time crime centers, which started in bigger cities, spread across the U.S. *NPR*. https://www.npr.org/2023/08/16/1194115202/real-time-crime-centers-which-started-in-bigger-cities-spread-across-the-u-s

Jeremy Binckes. (2023, November 15). Yes, people are still using 'password' for their password. *Msn.Com*. https://www.msn.com/en-us/money/other/yes-people-are-still-using-password-for-their-password/ar-AA1jXZSW

Jiazhen Ji, Huan Wang, Yuge Huang, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, ShengChuan Zhang, Liujuan Cao, & Rongrong Ji. (2022). Privacy-Preserving Face Recognition with Learnable Privacy Budgets in Frequency Domain. *ECCV 2022: Computer Vision – ECCV 2022*, 475–491.

Jon Russell. (2017, September 4). Alibaba debuts 'smile to pay' facial recognition payments at KFC in China. *TechCrunch*. https://techcrunch.com/2017/09/03/alibaba-debuts-smile-to-pay/

Joy Buolamwini & Timnit Gebru. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline

Kashmir Hill. (2020, August 3). Wrongfully Accused by an Algorithm. *New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Kay L. Ritchie, Charlotte Cartledge, Bethany Growns, An Yan, Yuqing Wang, Kun Guo, Robin S. S. Kramer, Gary Edmond, Kristy A. Martire, Mehera San Roque, & David White. (2021, October 13). Public attitudes towards the use of automatic facial recognition technology in criminal justice systems around the world. *PLoS ONE*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8513835/

Ken Cabeen & Peter Grant. (n.d.). *Image Compression and the Discrete Cosine Transform*. College of the Redwoods. https://www.math.cuhk.edu.hk/~lmlui/dct.pdf

Liyue Fan. (2018). *Image Pixelization with Differential Privacy*. https://link.springer.com/chapter/10.1007/978-3-319-95729-6_10

M.A.P. Chamikara, P. Bertok, I. Khalil, D. Liu, & S. Camtepe. (2020). Privacy Preserving Face Recognition Utilizing Differential Privacy. *Computers & Security*, *97*. https://www.sciencedirect.com/science/article/pii/S0167404820302273

Mei Wang & Weihong Deng. (n.d.). *Ethnicity Aware Training Datasets* [dataset]. http://www.whdeng.cn/RFW/Trainingdataste.html

Nick Galov. (2023, May 20). 20 Facial Recognition Statistics to Scan Through in 2023. *Web Tribunal*. https://webtribunal.net/blog/facial-recognition-statistics/

Oloid Desk. (2023, November 6). Facial Authentication Revolution: 15 Industries Embracing the Future of Security. *OLOID*. https://www.oloid.ai/blog/facial-authentication-revolution/

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, & Andrew Zisserman. (2018). *VGGFace2: A dataset for recognising faces across pose and age*. https://www.robots.ox.ac.uk/~vgg/publications/2018/Cao18/cao18.pdf