Limiting Privacy Incursion from Facial Recognition through De-identifying Face Images in the Public Domain

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

John Phillips

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Richard D. Jacques, Department of Engineering and Society

Introduction

Facial recognition technology has seen profound advancements in recent years. This evolution has allowed this technology to be developed into tools with severe implications regarding privacy. These privacy concerns are juxtaposed against this technology’s massive potential for good. This capability for good was seen in China where “police were able to identify and apprehend a criminal at a music concert attended by 60,000 people” (Walsh, 2022). In New Delhi, facial recognition “reunited nearly 3,000 children with their parents” (Walsh, 2022). In the United States, a Florida man was wrongly accused of vehicular homicide and facial recognition software was used to exonerate him. After spending “hundreds of hours” looking for the sole witness to the accident, this technology located the witness “within two seconds…at some club in Tampa” (Hill, 2022). This use of facial recognition differs from the previous two in a fundamental way–the dataset used to “match” faces. The criminal at the music concert had a known mugshot on file. The children reunited with their parents were matched against photos submitted for the specific intention at hand. However, the database of images used to identify the witness that exonerated the Florida man is owned and operated by a New York-based startup–Clearwater AI. This company has created a database “of people’s faces from across the internet, such as employment sites, news sites, educational sites, and social networks including Facebook, YouTube, Twitter, Instagram, and even Venmo ” (Hill, 2020). The witness was found using a photo posted on Facebook–a photo collected and used without their consent. This use of facial recognition highlights both the unprecedented capabilities of this technology as well as the danger to privacy these advancements pose.

The emergence of companies like Clearwater AI illustrates how publicly available photos are a powerful tool for amassing a database of individuals who may not have a face photo available by other means. The capabilities these datasets offer are incomprehensible, and come at the expense of individual privacy. This can be seen through Clearwater AI; this company has taken to scraping pictures–storing photos and associated information–from the public domain. I assert that privacy in the digital age, specifically protection against facial recognition technology, must include protecting personal images used to fuel facial recognition algorithms. The purpose of this paper is to explore technologies used to obfuscate people's identities and other sensitive personal information from images while maximizing quality and photo-realistic nature. Ideally, this technology would prevent the massive scraping of the public domain by making personal images unidentifiable to facial recognition algorithms. Every individual has different privacy needs and considerations which must be balanced against the retention of quality and the photo-realistic nature of the original image. It is this variability in user needs that has not been discussed in current facial recognition research. This paper will specifically investigate this user determined tradeoff between preventing facial recognition identification and the quality of the protected image.

Methodologies

In exploring technology that obfuscates people's identities and sensitive information from images while retaining the image's quality and photo-realistic nature, this paper considers two facets of an image. First, the performance of current facial recognition algorithms on the protected image. Second, the ability of a human to notice decreased image quality and degradation of the image's photo-realistic nature. These two parameters must be considered to correctly assess the tradeoff between protection against facial recognition identification and degradation of image quality to a user.

The first parameter under consideration, the ability of facial recognition algorithms to correctly identify facial images, is a purely quantitative measurement. There are two metrics needed to assess the performance of facial recognition algorithms against a protected image. The first revolves around the ability of a facial recognition algorithm to compare the original image to the protected image. This works by feeding the altered image into a facial recognition algorithm along with the original image. The algorithm provides a “score of confidence that indicates the probability that two faces belong to the same person” (Xue et al., 2023). The greater the confidence score, the higher probability that the two input images belonged to the same person. The second type of testing regards the protected image’s ability to protect an individual’s identity from a commercial network. These networks are large facial recognition networks; they use advanced deep neural networks and are trained on large datasets. The two most widely used commercial networks are Microsoft Face API and Face Plus Plus API. To conduct a test, the protected image is fed into these networks which output potential identity matches along with their associated score of confidence. Both facial recognition testing methods are widely used in research and are well-developed and documented.

On the other hand, calculating the ability of a human user to notice changes in a protected image is far less measurable than that of a computer. The amount of change in an image from the obfuscation process is easily quantifiable, however, a user’s perceived amount of change is not. A user’s ability to recognize change is heavily influenced by the user's familiarity with the face image through common features or direct knowledge. Furthermore, exposure to the original image would allow an observer to provide an elevated level of scrutiny regardless of familiarity with the face image. These factors complicate the process of estimating a user’s perceived change. Additionally, this testing would be subject to typical constraints imposed by surveying users. These include “the use of non-random sampling techniques…the reliance on biased perceptions…the problem of common method bias, and the problem of correlated omitted variables” (Speklé and Widener, 2018). Considering these limitations, determining a user's perceived change in an image from facial recognition privacy schemes is a hard measurement to quantify.

Privacy Implications of Facial Recognition

Facial recognition technology is a novel use of artificial intelligence. This technology employs machine learning models: algorithms that find patterns or make predictions based on a set of data. There have been a multitude of facial recognition algorithms designed, each with unique attributes. However, these algorithms can only be as powerful as the data they are trained with. The importance of the data used to train these models has led to the widespread collection of images, a process known as web scraping, from the public domain. Publicly available photos are a powerful tool for amassing a database of individuals who may not have a face photo available by other means. One example of this widespread data collection from the public domain is Clearview AI. This company has amassed "2.8-billion face photos…creating a search engine for any face image hosted on the public Internet" (Marks, 2021). In comparison, the FBI has access to 411 million face images. Additionally, the FBI has amassed its database through government-provided images, such as mug shots and driver's license photos. In contrast, Clearwater AI has taken to collecting images from every corner of the internet–without the consent of the individual. This example of web scraping face pictures highlights a pressing need to protect individual privacy in the images used to fuel facial recognition algorithms.

Current De-identification Techniques

Various techniques have been proposed to enhance individual privacy on the internet through de-identifying face images. The de-identification process aims to create a deliberate alteration to the image that is small but capable of deceiving facial recognition algorithms while simultaneously avoiding any noticeable change that can be detected by humans. Traditional obfuscation-based methods usually obscure sensitive information by blurring, pixelating, or masking an image" (Xue et al., 2023). However, these techniques can be "easily and accurately detected by deep learning models" (Xue et al., 2023). Therefore, new techniques and methods were created to de-identify images specifically against facial recognition algorithms, specifically ones using deep learning models. These methods include "differential privacy, GAN-based inpainting, and adversarial examples" (Xue et al., 2023). Techniques that utilize differential privacy (DP) can offer assurances of privacy that can be proven mathematically, but they can also result in images of reduced quality. To avoid this issue, GAN-based methods for inpainting have been suggested. These techniques generate content to conceal the identity of an image without compromising the quality of the initial image. However, GAN-based methods incorporate latent noise in a vague direction, without considering the unique characteristics of the face image. Considering the advancement of facial recognition and the computing techniques powering these algorithms, specifically large-scale neural networks, adversarial examples (AE)-based protection methods hold significant promise. The earliest research on AE showed that "a small perturbation could have a considerable and negative impact on the accuracy of deep neural networks" (Szegedy et al., 2014). One research team continuing this research is based out of the University of Technology Sydney. This group is at the forefront of creating AE-based protection methods; they have published *Hiding Private Information in Images From AI* (Xue et al., 2018) and more recently *Face image de-identification by feature space adversarial perturbation* (Xue et al., 2023). This recent publication proposed a novel face image de-identifying framework. The highlight of this framework is the addition of adversarial perturbations on the feature level of images rather than indiscriminately adding noise to an image. This serves to anonymize face identity information against automated recognition while maintaining the quality of an image. This technique provides improved image quality while maintaining protection against facial recognition. However, even this advanced (AE)-based protection method does not account for differing user needs in terms of balancing image quality with protection.

One proposed solution that allows for differing levels of protection was published by a research team from Fudan University in Shanghai. This team proposed a novel reversible face-privacy preservation scheme. This framework accounts for three types of users: non-authorized or illegal users, low-level users, and authorized users. “Illegal users obtain no identity information from the [images], while low-privileged users can use protected images for computer vision tasks … authorized recipient[s] can recover the original facial images with high quality” (You et al., 2020). This research is a step in the right direction, as different users require various levels of protection. However, this proposal has little application for protecting users’ face photos in the public domain, especially regarding social media data. This is because this scheme requires a well-defined user space and lacks preservation of photo quality for low-level users.

Recommendations for Future Work

The optimal solution for preventing massive scraping of the public domain by rendering this data unusable to facial recognition algorithms requires a variable protection algorithm. This algorithm needs to be able to provide maximum protection against facial recognition at the cost of photo quality while also having the capability to produce a high photo quality photo at the cost of protection. This would allow a user to determine the tradeoff between protection and photo quality based on their personal needs. The capability to adapt to different user needs would allow for widespread acceptance of the practice of de-identifying face images. This, in turn, would minimize public domain web scraping for the purpose of facial recognition.

Conclusion

The profound advancements in facial recognition technology in recent years have evolved the technology into a tool that has severe implications regarding privacy. These privacy concerns stem from the widespread web scraping of face images from the public domain. Web scraping is conducted both to train these facial recognition algorithms as well as provide identifying information associated with an input face photo. The capabilities these datasets offer are unprecedented, however, it comes at the expense of individual privacy. This paper asserts that digital privacy, specifically protection against facial recognition technology, stems from protecting personal images used to fuel facial recognition algorithms.

Technologies currently exist to de-identify face images. These include traditional obfuscation-based methods as well as newer AE-based protection methods. Obfuscation-based methods have proven to be ineffective due to advancements in computing technologies, specifically facial recognition algorithms employing deep learning models. On the other hand, AE-based methods have proven effective against these deep learning models. AE-based methods incur minimal changes to the image while providing extensive de-identification. The most recent advancements in these methods proposed in the article *Face image de-identification by feature space adversarial perturbation* (Xue et al., 2023) focus the AE algorithm on facial features, rather than indiscriminately applying changes to the entire image. This advancement has led to both increased protection and image quality. However, this technology will not be able to minimize web scraping of the public domain. This limitation stems from the lack of variability in the protection algorithm, leaving it unable to meet the needs of a wide range of users. Every individual has different privacy needs and considerations which must be balanced against retention of quality and photo-realistic nature of the original image. This variability in user need has not been discussed in current facial recognition research and continues to prevent the widespread acceptance of de-identifying face images uploaded to the internet.

References

Andrade, N. N. G. de, Martin, A., & Monteleone, S. (2013a). “All the better to see you with, my dear”: Facial recognition and privacy in online social networks. *IEEE Security & Privacy*, *11*(3), 21–28.

Andrade, N. N. G. de, Martin, A., & Monteleone, S. (2013b). “All the better to see you with, my dear”: Facial recognition and privacy in online social networks. *IEEE Security & Privacy*, *11*(3), 21–28.

Chi Liu, Tianqing Zhu, Jun Zhang, & Wanlei Zhou. (2023a). Privacy Intelligence: A Survey on Image Privacy in Online Social Networks. *ACM Computing Surveys*, *55*(8), 1–35.

Chi Liu, Tianqing Zhu, Jun Zhang, & Wanlei Zhou. (2023b). Privacy Intelligence: A Survey on Image Privacy in Online Social Networks. *ACM Computing Surveys*, *55*(8), 1–35.

Goswami, G., Agarwal, A., Ratha, N., Singh, R., & Vatsa, M. (2019). Detecting and Mitigating Adversarial Perturbations for Robust Face Recognition. *International Journal of Computer Vision*, *127*(6/7), 719–742.

Hill, K. (2020, January 18). The Secretive Company That Might End Privacy as We Know It. *The New York Times*.

Hill, K. (2022, September 18). Clearview AI, Used by Police to Find Criminals, Is Now in Public Defenders’ Hands. *The New York Times*.

Jagadeesha, N. (2022a). Facial Privacy Preservation using FGSM and Universal Perturbation attacks. *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, *1*, 46–52.

Jagadeesha, N. (2022b). Facial Privacy Preservation using FGSM and Universal Perturbation attacks. *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, *1*, 46–52.

Johnston, L. (2020). Ninth Circuit Rejects LinkedIn's Efforts to Block Web-Scraping of Member Public Profiles. *Computer & Internet Lawyer*, *37*(4), 5–7.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks* (arXiv:1312.6199). arXiv.

Walsh, T. (2022). The Troubling Future for Facial Recognition Software: Considering the myriad perspectives of facial recognition technology. *Communications of the ACM*, *65*(3), 35–36.

Xue, H., Liu, B., Din, M., Song, L., & Zhu, T. (2020). Hiding Private Information in Images From AI. *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 1–6.

Xue, H., Liu, B., Yuan, X., Ding, M., & Zhu, T. (2023). Face image de-identification by feature space adversarial perturbation. *Concurrency & Computation: Practice & Experience*, *35*(5), 1–13.

You, Z., Li, S., Qian, Z., & Zhang, X. (2021). Reversible Privacy-Preserving Recognition. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.