Adversarial Facial Obfuscation against Unauthorized Face Recognition

Jingbo Hao* and Yang Tao

College of Artificial Intelligence, Nanchang Institute of Science and Technology, 998 Gezaoshan Ave., Honggutan District, Nanchang, China.
Email: hjb@nudt.edu.cn

Abstract. To protect individual privacy from unauthorized face recognition based on DNN models, adversarial facial obfuscation tries to generate an adversarial image with a feature vector differing markedly from the original image in the embedding space and keep perceptually similar between the two images simultaneously. This paper makes a brief survey of adversarial facial obfuscation. The preliminary theory about facial obfuscation is introduced first. With regard to adversarial facial obfuscation, the most important implementation factors consisting of transferability, perceptibility and compression resistance are also presented.

1. Introduction

The optimization of face recognition systems goes further with the rapid development of deep learning technology and deep neural networks (DNNs). Nowadays a perfect face recognition service behaves quite well even under some abnormal conditions such as makeup, masks and face spoofing [1]. However the misuse of such technology has been emerging. For example a private company named Clearview AI was reported to scrape more than 3 billion face images from Facebook, YouTube, Venmo and millions of other websites and provide face recognition service to many law enforcement agencies and several companies [2]. Moreover, social media companies or online photo application providers can also collect user images and use these images for face recognition and activity analysis without explicit prompting.

To protect individual privacy from unauthorized face recognition, facial obfuscation (i.e., de-identification or anonymization) techniques aim at providing an acceptable trade-off between protection of facial identities and preservation of non-sensitive utility information in obfuscated images [3]. The most classical paradigm for provable facial obfuscation is k-Same de-identification which guarantees a recognition rate lower than 1/k [4]. Current k-Same solutions, such as k-Same-Pixel [4], k-Same-Eigen [4], k-Same-Select [7], k-Same-M [8], k-Same-Furthest [9] and k-Same-Net [10], can deliver passable privacy protection. In recent years, several deficiencies of k-Same de-identification (e.g., poor visual quality and susceptibilities to background knowledge or composition attacks) are listed in [3] and may reduce k-Same solutions’ applicability. Fan [11, 12] has applied differential privacy in facial obfuscation to provide rigorous privacy protection, but the obfuscated images are not photo-realistic and cannot eliminate the possibility of information leakage. To address the deficiencies Croft et al [13] developed the first framework applying differential privacy for facial obfuscation on generative models to prevent facial identification while maintaining utility and visual quality.
However, on many occasions users prefer to conceal their identities from unauthorized face recognition and not affect common face detection or human perceptibility at the same time, which makes former obfuscation solutions not so suitable. Due to the black-box reality of unauthorized face recognition systems, an ideal facial obfuscation solution for that purpose should be applied in a pure black-box manner (i.e. nothing about the DNN model for a face recognition system is known, such as its parameters, training data or even outputs) and made available transferably for other different systems. Nowadays the research about adversarial attacks on deep learning models in computer vision has become a hotspot [14]. The reason is that DNNs are often vulnerable to adversarial examples [15] which are usually imperceptible to human observers by adding small perturbations and can mislead a vulnerable DNN model to produce wrong predictions. The particular property of adversarial examples is a great inspiration for researchers to confront unauthorized face recognition and some instructive attempts on adversarial facial obfuscation [16-23] have been made.

This paper focuses on an up-to-date survey of adversarial facial obfuscation against unauthorized face recognition. The preliminary theory about facial obfuscation is briefly introduced in Section 2. With regard to adversarial facial obfuscation, the most important implementation factors consisting of transferability, perceptibility and compression resistance are discussed in Section 4. Finally the conclusions are drawn.

2. Preliminary Theory

2.1. Basic Definitions

A set of formal definitions about the problem of facial de-identification have been given in [4, 5] and are slightly revised as follows.

**Definition 2.1 (Face Image).** A face image is a 3D matrix $I$ of $m$ columns, $n$ rows, and $c$ channels. Each cell in $I$ stores a color coding for a pixel, ranging from 0 to 255 inclusively. A face image contains a normalized image of only one person’s face.

**Definition 2.2 (Face Set).** A face set is a set of $M$ face images: $\{\Gamma_i; i = 1, \ldots, M\}$.

**Definition 2.3 (Person-Specific Face Set).** Let $H$ be a face set of $M$ face images $\{\Gamma_1, \ldots, \Gamma_M\}$. $H$ is said to be person-specific if and only if each $\Gamma \in H$ relates to only one person, and no two images $\Gamma_i \in H$ and $\Gamma_j \in H$ where $i \neq j$ relate to the same person.

**Definition 2.4 (Face De-Identification).** Let $H$ and $H_d$ be face sets, $\Gamma \in H$, $\Gamma_d \in H_d$, $f: H \rightarrow H_d$ be a function that attempts to conceal the identity of the subject of the original face image, and $\Gamma_d = f(\Gamma)$ but $\Gamma \neq \Gamma_d$, then $f$ is called face de-identification (function) and $\Gamma_d$ is a de-identified image.

**Definition 2.5 (Effective De-Identification).** Let $H$ be a person-specific face set, $H_d$ be a face set, $f: H \rightarrow H_d$ be a face de-identification function, $g: H_d \rightarrow H$ be a face identification (recognition) relation, and $C$ be a provable claim about the ability of $f$ to restrict face identification by $g$, then $f$ provides effective de-identification with respect to $C$. If $f_1$ and $f_2$ are effective with respect to the same $C$, then $f_1$ and $f_2$ are equally effective with respect to $C$.

2.2. Traditional Privacy Metrics

The goal of privacy metrics is to measure the degree of privacy possessed by a system user and the amount of protection offered by privacy-enhancing techniques [6]. The following definitions mainly derive from [3, 4, 11].

**Definition 2.6 (k-Anonymity).** Given a person-specific face set $H$, a set of de-identified face images $H_d$ in which duplicates are maintained, $|H| > 1$, $|H| = |H_d|$, a de-identification function $f: H \rightarrow H_d$, a face identification relation $g: H_d \rightarrow H$. If for each $\Gamma \in H$ there exists $\Gamma_d \in H_d$, where $\Gamma_d = f(\Gamma)$ and for each $\Gamma_d \in H_d$, $|g(\Gamma_d)| = \Gamma \geq k$, then $H_d$ adheres to $k$-anonymity and $H_d$ is $k$-anonymized over $H$.

**Definition 2.7 (k-Same).** Given a person-specific face set $H$, and a face set $H_d$ which is $k$-anonymized over $H$ using a preserved face de-identification function $f: H \rightarrow H_d$, if $f$ is effective with respect to the claim: “given any face image $\Gamma_i \neq f(\Gamma)$ for $\Gamma \in H$, there cannot exist any face recognition
software for which the subject of $\Gamma_s$ can be correctly recognized as $\Gamma$ with better than $1/k$ probability,” then $f$ is a $k$-Same de-identification function and $H_d$ is a $k$-Same de-identification. A de-identification function $f$ is preserved or preservative if it minimizes the information loss resulting from de-identification with respect to a monotonic metric $loss(\Gamma, f(\Gamma))$ among a set of equally effective de-identification functions.

Definition 2.8 ($m$-Neighborhood). Two images $I_1$ and $I_2$ are neighboring images if they have the same dimension and they differ by at most $m$ pixels.

Definition 2.9 ($\epsilon$-Differential Privacy). A randomized mechanism $A$ gives $\epsilon$-differential privacy if for any neighboring images $I_1$, $I_2$, and any possible output $L \in \text{Range}(A)$, the following claim shown in Formula 1 is met:

$$\Pr[A(I_1) = L] \leq e^\epsilon \times \Pr[A(I_2) = L] \tag{1}$$

Definition 2.10 (Generative Differential Privacy). Let the generative model representation of an image be a vector $X \in \mathbb{R}^n$, $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be a distance function for a pair of image representations, then a randomized mechanism $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ gives generative differential privacy if the following claim shown in Formula 2 is met:

$$\Pr[A(X_1) = X_d] \leq e^{d(X_1,X_2)} \times \Pr[A(X_2) = X_d] \quad \forall X_1, X_2, X_d \in \mathbb{R}^n \tag{2}$$

2.3. Adversarial Privacy Metrics

Two definitions about transferability-based privacy metrics for adversarial facial obfuscation are tentatively given below.

Definition 2.11 (Transferable De-Identification). Let $H$ be a person-specific face set, $H_d$ be a face set, $f: H \rightarrow H_d$ be a face de-identification function, $g: H_d \rightarrow H$ be a face identification function, $R$ is a set of face identification functions, $g \in R$ and $|R|>1$, $C$ is a claim about the ability of $f$ to restrict face identification by a certain face identification function. If $C$ is satisfied on $g$, and accordingly $C$ are satisfied on each $g \in R$, then $f$ provides transferable de-identification from $g$ to $R$ with respect to $C$.

Definition 2.12 (Adversarial Privacy). Let $H$ be a person-specific face set, $H_d$ be a face set, $f: H \rightarrow H_d$ be a face de-identification function, $g: H_d \rightarrow H$ be a face identification function, $R$ is a set of face identification functions, $g \in R$ and $|R|>1$, $C$ is a claim about the ability of $f$ to restrict face identification, $f$ provides transferable de-identification from $g$ to $R$ with respect to $C$. If $C$ is not satisfied on human observers, then $f$ provides adversarial privacy with respect to $C$.

3. Adversarial Facial Obfuscation

3.1. Transferability

Due to the black-box nature of unauthorized face recognition systems, practical adversarial facial obfuscation should be applied in a black-box manner. Black-box attacks hinge on the property of transferability enabling adversarial examples generated on known substitute models to transfer the attacking effects to unknown target models in a certain extent. To enhance the transferability for adversarial facial obfuscation the following schemes have been utilized:

1. Perturbation amplification. Amplification increases transferability at the expense of higher perturbation and consequently human perceptibility [17].
2. Expectation over transformation (EOT). EOT is able to produce robust adversarial examples that are simultaneously adversarial over an entire distribution of transformations [24]. The algorithm is well applied to adversarial facial obfuscation in [21].
3. Ensemble of substitute models. If an adversarial image remains adversarial for multiple substitute models, it is more likely to transfer to other unknown models [25]. This scheme is widely utilized in adversarial facial obfuscation solutions [20-23].
4. Robustness improvement of substitute models. It has been shown that an adversarial example’s ability to transfer between models depends on the robustness of the feature extractor used to
generate it [26]. This can be done by applying adversarial training to train the substitute model on perturbed data to make it more robust [18].

3.2. Perceptibility
Adversarial facial obfuscation tries to generate an adversarial image whose feature vector lies far away from the original image in the embedding space and minimize the perceptual similarity loss between the original and adversarial images simultaneously [23]. Zhang et al [27] proposed an effective perceptual metric - learned perceptual image patch similarity (LPIPS) which measures how similar two images are in a way coinciding with human observation. LPIPS correlates well with human judgements across different adversarial perturbations [28] and consequently is fit to compute perceptual similarity loss for adversarial facial obfuscation [23].

3.3. Compression Resistance
To counter against the threat of adversarial attacks, a simple but valid defence approach is JPEG compression which can remove high frequency components within an image and hence attenuate additive perturbations [29] as indeed a feature distillation process [30]. However, experimental results show that JPEG compression does not degrade adversarial facial obfuscation distinctly [17, 18, 23].

4. Conclusions
To protect individual privacy from unauthorized face recognition based on DNN models, adversarial facial obfuscation solutions emerged. With regard to adversarial facial obfuscation, the preliminary theory is briefly introduced and the most important implementation factors consisting of transferability, perceptibility and compression resistance are discussed here. Along with the continuous improvement of deep learning technology, more robust face recognition systems against adversarial perturbations will come into service and may be employed at no allowance. To counter this challenge, the research on adversarial facial obfuscation still possesses great development potential.

Acknowledgments
This work was partially supported by the Introduced Talent Research Start-up Fund of Nanchang Institute of Science and Technology under Grant No. NGRCZX-20-12. The authors are also grateful to those resource providers for their selfless help.

References
[1] George A and Marcel S 2020 Can your face detector do anti-spoofing? Face presentation attack detection with a multi-channel face detector arXiv Preprint arXiv:2006.16836
[2] Hill K 2020 The secretive company that might end privacy as we know it The New York Times 2020-01-18
[3] Croft WL, Sack JR and Shi W 2021 Obfuscation of images via differential privacy: from facial images to general images arXiv Preprint arXiv:2102.11072
[4] Newton EM, Sweeney L and Malin B 2005 Preserving privacy by de-identifying face images IEEE Transactions on Knowledge and Data Engineering vol 17 no 2 pp 232-243
[5] Li T and Lin L 2019 AnonymousNet: natural face de-identification with measurable privacy Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition Workshops pp 56-65
[6] Wagner I and Eckhoff D 2018 Technical privacy metrics: a systematic survey ACM Computing Surveys vol 51 no 3 article 57
[7] Gross R, Airoldi E, Malin B and Sweeney L 2005 Integrating utility into face de-identification Proc. 5th International Workshop on Privacy Enhancing Technologies pp 227-242
[8] Gross R, Sweeney L, Torre F and Baker S 2006 Model-based face de-identification Proc. 2006 IEEE Conf. on Computer Vision and Pattern Recognition Workshops
[9] Meng L and Sun Z 2014 Face de-identification with perfect privacy protection Proc. 37th International Conv. on Information and Communication Technology, Electronics and Microelectronics pp 1234-1239

[10] Meden B, Emersic Z, Struc V and Peer P 2017 k-Same-Net: neural-network-based face de-identification Proc. 2017 International Conf. and Workshop on Bioinspired Intelligence

[11] Fan L 2018 Image pixelization with differential privacy Proc. 32nd IFIP Annual Conf. on Data and Applications Security and Privacy pp 148-162

[12] Fan L 2019 Practical image obfuscation with provable privacy Proc. 2019 IEEE International Conf. on Multimedia and Expo pp 784-789

[13] Croft WL, Sack JR and Shi W 2019 Differentially private obfuscation of facial images Proc. 3rd IFIP International Cross-Domain Conf. for Machine Learning and Knowledge Extraction pp 229-249

[14] Akhtar N and Mian A 2018 Threat of adversarial attacks on deep learning in computer vision: a survey IEEE Access vol 6 pp 14410-14430

[15] Goodfellow IJ, Shlens J and Szegedy C 2015 Explaining and harnessing adversarial examples Proc. 3rd International Conf. on Learning Representations arXiv Preprint arXiv:1412.6572

[16] Oh SJ, Fritz M and Schiele B 2017 Adversarial image perturbation for privacy protection: a game theory perspective Proc. 2017 IEEE International Conf. on Computer Vision pp 1491-1500

[17] Chandrasekaran V, Gao C, Tang B, Fawaz K, Jha S and Banerjee S 2020 Face-Off: adversarial face obfuscation arXiv Preprint arXiv:2003.08861

[18] Shan S, Wenger E, Zhang J, Li H, Zheng H and Zhao BY 2020 Fawkes: protecting privacy against unauthorized deep learning models Proc. 29th USENIX Security Symp. pp 1589-1604

[19] Yang X, Dong Y, Pang T, Zhu J and Su H 2020 Towards privacy protection by generating adversarial identity masks arXiv Preprint arXiv:2003.06814

[20] Zhang J, Sang J, Zhao X, Huang X, Sun Y and Hu Y 2020 Adversarial privacy-preserving filter Proc. 28th ACM International Conf. on Multimedia pp 1423-1431

[21] Evtimov I, Sturmfels P and Kohno T 2020 FoggySight: a scheme for facial lookup privacy arXiv Preprint arXiv:2012.08588

[22] Rajabi A, Bobba RB, Rosulek M, Wright CV and Feng W 2021 On the (im)practicality of adversarial perturbation for image privacy Proceedings on Privacy Enhancing Technologies vol 2021 no 1 pp 85–106

[23] Cherepanova V, Goldblum M, Foley H, Duan S, Dickerson J, Taylor G and Goldstein T 2021 LowKey: leveraging adversarial attacks to protect social media users from facial recognition Proc. 9th International Conf. on Learning Representations arXiv Preprint arXiv:2101.07922

[24] Athalye A, Engstrom L, Ilyas A and Kwok K 2018 Synthesizing robust adversarial examples Proc. 35th International Conf. on Machine Learning pp 284-293

[25] Liu Y, Chen X, Liu C and Song D 2017 Delving into transferable adversarial examples and black-box attacks Proc. 5th International Conf. on Learning Representations arXiv Preprint arXiv:1611.02770

[26] Demontis A, Melis M, Pintor M, Jagielski M, Biggio B, Oprea A, Nita-Rotaru C and Roli F 2019 Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks Proc. 28th USENIX Security Symp. pp 321–338

[27] Zhang R, Isola P, Efros AA, Shechtman E and Wang O 2018 The unreasonable effectiveness of deep features as a perceptual metric Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition pp 586-595

[28] Laidlaw C, Singla S and Feizi S 2021 Perceptual adversarial robustness: defense against unseen threat models Proc. 9th International Conf. on Learning Representations arXiv Preprint arXiv:2006.12655

[29] Das N, Shanbhogue M, Chen ST, Hohman F, Chen L, Kounavis ME and Chau DH 2017 Keeping the bad guys out: protecting and vaccinating deep learning with JPEG compression arXiv Preprint arXiv:1705.02900
[30] Liu Z, Liu Q, Liu T, Xu N, Lin X, Wang Y and Wen W 2019 Feature distillation: DNN-oriented JPEG compression against adversarial examples Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition pp 860-868