AnonymousNet: Natural Face De-Identification with Measurable Privacy

Tao Li and Lei Lin

Purdue University
University of Rochester

CVPR LONG BEACH CALIFORNIA
June 16-20, 2019

Outline

- Motivation & Background
- Our Approach: The AnonymousNet
- Experiments
- Discussion & Future Works

Privacy v.s. Usability

Face Obfuscation

(a) Original (b) Blurring (c) Pixelation (d) Masking (e) Abstract (f) Portrait (g) Cartoon (h) Inpainting

Face Obfuscation

DeepFake

Nirkin et al. FG'18

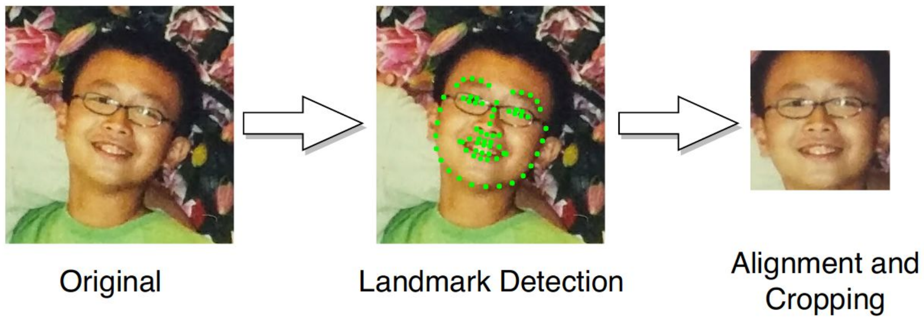
Sun et al. CVPR'18

Unanswered Questions

- Is it private now?
- How private is it?
- Can it be more private/usable?
- Why?

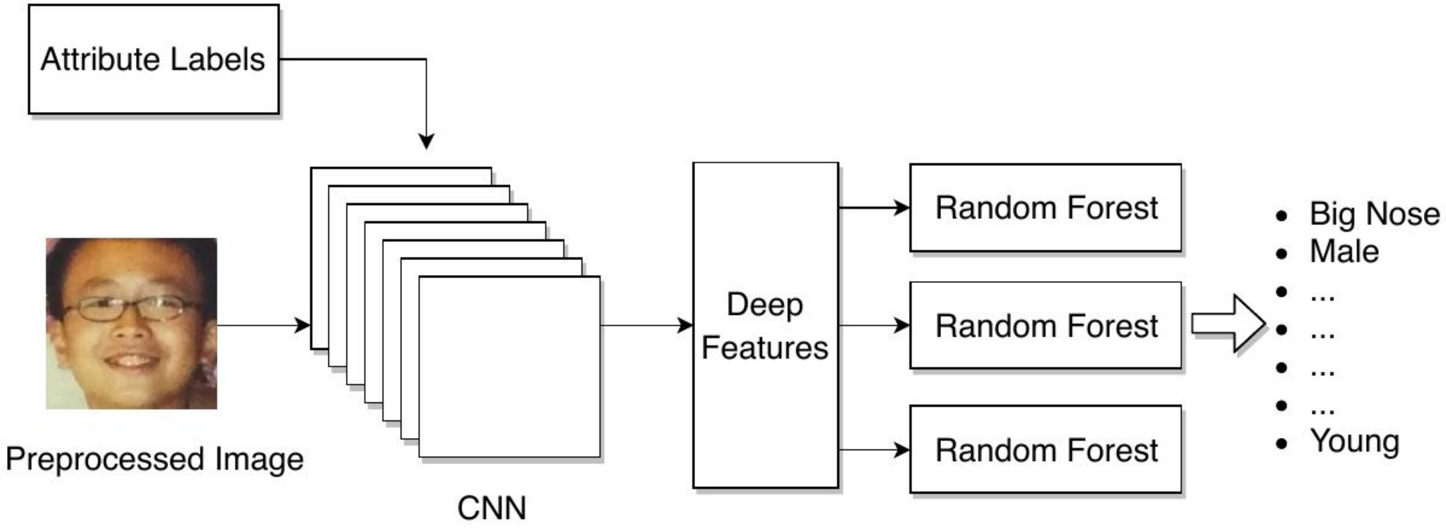
AnonymousNet: A Natural and Principled Way for Face Obfuscation

Stage-I: Facial Attribute Prediction Using CNN

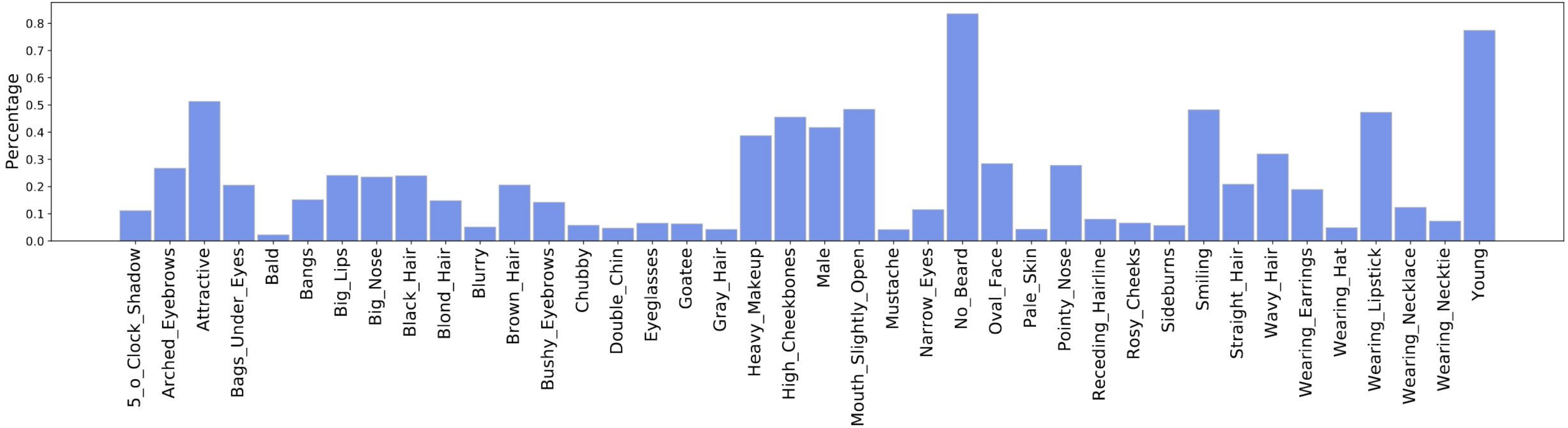


Preprocessing using a Deep Alignment Network (Kowalski et al. CVPR'17)

Stage-I: Facial Attribute Prediction Using CNN



Stage-II: Privacy-Aware Facial Semantic Obfuscation



Using CeleA dataset (Liu et al. ICCV'15) as an example.

Stage-II: Privacy-Aware Facial Semantic Obfuscation

t-Closeness Adversaries sometimes have knowledge of the global distribution of sensitive attributes, for example, the distributions of facial attributes are easy to obtain (see Figure 6). To prevent privacy disclosure by an adversary with such knowledge, [24] introduced t-closeness, which updates k-anonymity with correspondence to the distribution of sensitive values, requiring that the distribution S_E of sensitive values in any equivalence class E must be close to their distribution S in the entire database, i.e.,

$$\forall E : d(S, S_E) \leq t \tag{5}$$

where $d(S, S_E)$ is the distance between distribution S and S_E measured by the Earth Mover Distance [47] and t is the privacy threshold at which $d(S, S_E)$ should not exceed.

Algorithm 1: The PPAS algorithm.

```
Result: Attribute set 𝔸″.
Attribute set 𝔸 ← {E_1, ..., E_n};
Attribute set 𝔸′ ← ∅ ;
Size N ← ||𝔸|| ;
for i = 1, ..., N do
    if d(S, S_{E_i}) ≤ t then
        Add attribute E_i to 𝔸′ ;
    else
        Add attribute ¬E_i to 𝔸′ ;
    end
end
return 𝔸″ ← Perturbation(𝔸′, ε) ;
```

Privacy-Preserving Attribute Selection

Stage-III: Natural Face Generation Using GAN

After obtaining facial attributes that satisfies privacy constraints computed from previous steps, we train a Generative Adversarial Network (GAN) for face attributes translation, which is designed as two players, D and G, playing a minmax game with adversarial loss:

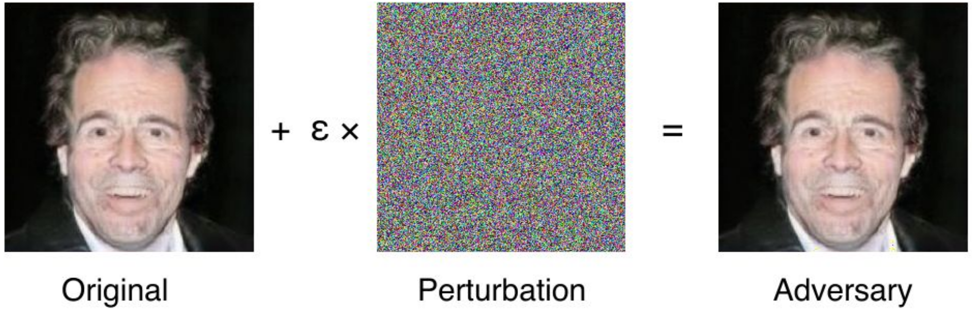
$$L_{adv} = \mathbb{E}[\log(D(\mathbf{x}))] + \mathbb{E}[\log(1 - D(G(\mathbf{x})))] \qquad (1)$$

where generator G is trained to fool discriminator D, who tries to distinguish real images from adversarial ones.

Choi et al. CVPR'18

Generated Examples.

Stage-IV: Adversarial Perturbation against Adversaries



Experimental Results

Comparison

(a) Original (b) Blurring (c) Pixelation (d) Ours

Summary

- We proposed the AnonymousNet for natural face de-identification.

- The framework encompasses four stages: facial feature prediction, semantic-based facial attribute obfuscation guided by privacy metrics, photo-realistic and de-identified face generation, and adversarial perturbation.

- Privacy is preserved in a natural and principled manner.

Next Steps

- A formally definition of ε-Differential Privacy for facial images.

- Principled and end-to-end models for privacy preservation.

- Extended frameworks for sequential domains.

Thank you!

Poster #134 | @Tao_CS

The paper is available on: https://arxiv.org/abs/1904.12620