From Understanding to Controlling Privacy against Automatic Person Recognition in Social Media

Seong Joon Oh, Mario Fritz, Bernt Schiele
Max-Planck Institute for Informatics, Germany

goo.gl/hgvbNK

Background & Motivation

- Huge amount of shared personal photos online: 40 billion images on Instagram alone (2017).
- Computer vision works really well.
- Malicious entity could use computer vision technique to spy into users' privacy through uploaded photos.

Goal & Roadmap

Goal: Anonymise subjects in social media photos without visible artifacts.

1. Understand how well subjects can be identified in social media photos.
2. Measure how successfully common anonymisation techniques are.
3. Develop effective anonymisation techniques.

1. Person Recognition in Social Media, ICCV'15 [2]

Task: Who is this person among {A,B,…} ?

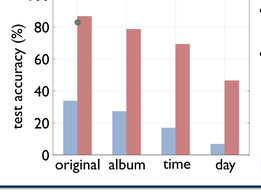
Image + head box

Dataset: PIPA [1,2]

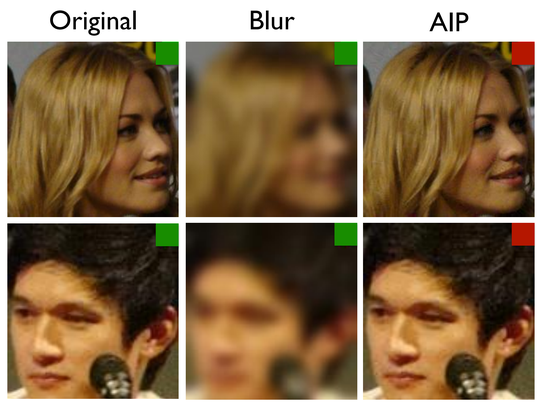
Method

- Train region specific feature extractors.
- Extract features & concatenate.
- Train identity classifiers.

Results



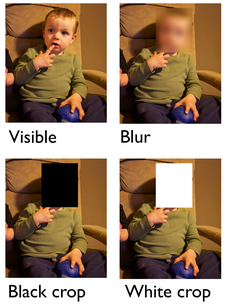
- 87% in Original split; 47% in Day split.
- Subjects can be identified via simple convnet technique.



- Blurring is unpleasant and ineffective.
- AIPs are invisible and effective, but only for known recogniser.
- How to make it work for unknown recognisers? – Game Theory.

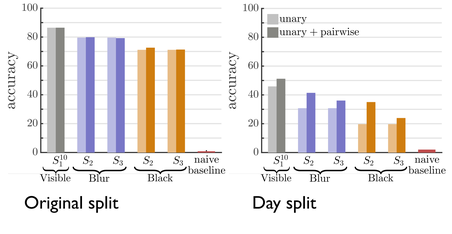
2. Defeating Common Anonymisation, ECCV'16 [3]

Anonymisation Scenarios



- S1: Fully visible.
- S2: Blur/ crop one head of interest.
- S3: Blur/ crop all the heads.

Results

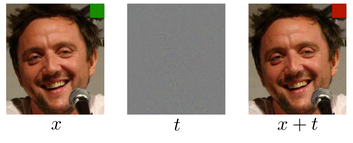


- Blurring or cropping is not effective.
- Even for the hardest setup (S3, Black crop, Day split), recognition rate is 12x the random chance.

3. AIP Anonymisation and Game Theory, ICCV'17 [4]

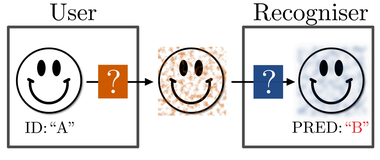
Adversarial Image Perturbations (AIP)

$$\max_t \mathcal{L}\left(f\left(x+t\right), y\right) \qquad \text{s.t. } ||t||_2 \leq \epsilon$$



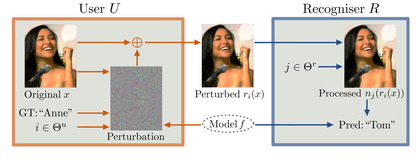
- Loss *maximisation* problem.
- Invisible to human eyes, but effectively fools machines.

Game Theoretical Framework



- Motivation: AIP may not work against a non-targetted machine.
- Game Theory: model uncertainty in the opponent's chosen mechanism.

A Case Study



- Evaluate against defense methods: Translate, Noise, Blur, Crop, Ensemble [5]. (Recogniser strategy)
- Robustify the vanilla AIP against each defense method. (User strategy)

Recogniser *R*

User *U*	Vanilla	Trans.	Noise	Blur	Crop	Ens.
Vanilla AIP	4.0	6.6	15.0	22.2	16.7	9.9
Trans. robust	2.5	2.3	11.6	18.5	7.2	4.9
Noise robust	5.8	7.6	4.6	23.6	16.6	9.1
Blur robust	0.4	0.8	8.6	5.8	3.1	1.4
Crop robust	2.6	2.2	11.8	18.1	3.4	4.3
Ens. robust	0.7	0.9	5.2	9.5	3.2	2.0

Table: Recognition success rates (%) when a strategy pair for *R* and *U* are played.

- Optimal *U* strategy: "Blur robust" 61% and "Ensemble robust" 39%.
- Guarantees at most **7.4%** recognition rate, independent of *R*'s action.

References

[1] Zhang et al. Beyond Frontal Faces: Improving Person Recognition Using Multiple Cues. CVPR'15
[2] Oh et al. Person Recognition in Personal Photo Collections. ICCV'15.
[3] Oh et al. Faceless Person Recognition; Privacy Implications in Social Media. ECCV'16.
[4] Oh et al. Adversarial Image Perturbation for Privacy Protection – A Game Theory Perspective. ICCV'17.
[5] Graese et al. Assessing Threat of Adversarial Examples on Deep Neural Networks. ICMLA'16.

Acknowledgement: Supported by German Research Foundation (DFG CRF 1223).