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

{joon,mfritz,schiele}@mpi-inf.mpg.de

Abstract

Growth of the internet and social media has spurred the sharing and dissemination of personal data at large scale. At the same time, recent developments in computer vision has enabled unseen effectiveness and efficiency in automated recognition. It is clear that visual data contains private information that can be mined, yet the privacy implications of sharing such data have been less studied in computer vision community. This extended abstract presents a line of research that begins with the study of person identification in social media photos and progresses towards effective computer vision technology for anonymisation.

1. Understanding Privacy

Person recognition in social media photos is a fledgling area in computer vision research; traditional focus has been face recognition and pedestrian re-identification, where vision algorithms have achieved impressive performances. Social media sets new challenges: non-frontal faces, severe occlusions, varying pose, etc. It is unclear how existing face recognition and re-identification techniques transfer to the social media setup. Relevant benchmarks and experimental setups have to be designed, and different methods need to be tried.

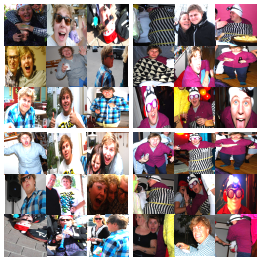

Figure 1. PIPA samples for identity X. *Original* (*Day*) split training and testing samples are shown in upper (left) and lower (right) halves.

This section reviews the first large-scale dataset and benchmark for person recognition in social media setup [4], and more realistic and challenging scenarios introduced by [1]. We describe the state of the art person recogniser [1].

Social media dataset: PIPA. The PIPA dataset [4] consists of social media photos on Flickr. It contains ∼40k images over ∼2k identities, and captures subjects appearing in diverse social groups and events. It is arguably the first large-scale social media dataset for person recognition.

The first setup Zhang *et al*. [4] considered evaluates the recognition system on the samples of identities from the same contexts (clothing, event, *etc*.) as the training samples, referred to as the *Original* split. Oh *et al*. [1] has introduced more challenging scenarios according to the domain shift between the training and test samples (*Album*, *Time*, and *Day*). Figure 1 visualises the *Original* and *Day* splits.

State of the art recogniser. Oh *et al*. [1] has proposed a simple yet effective framework. It trains convnet features for five body regions; it then fits a person classifier on the concatenated features using identity-specific training samples. See figure 2 for visualisation. [3] has extended the AlexNet based architecture in [1] to VGG, GoogleNet, and ResNet152, leading to better performances.

Figure 2. Person recogniser by [1].

Results. Figure 3 shows the performance of the recognition system. In the *Original* and *Day* splits, the system reaches 86.8% and 46.6% accuracy compared to the random baselines 0.8% and 2.0%, respectively, showing that the identity information can be successfully mined. For more details, see [1].

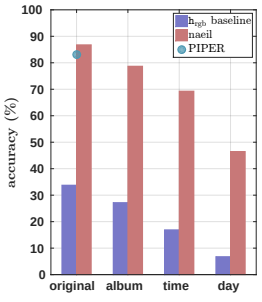

Figure 3. Recognition accuracy on different setups.

2. Controlling Privacy

It is often users' interest to anonymise subjects in images. Users typically employ face blurring or cropping out to ob-

fuscate identities in images (figure 4). This is often unpleasant and moreover not as effective as one would expect, since modern recognition systems can adapt to anonymisation patterns and crawl cues via social media metadata [2].

We advocate adversarial image perturbations (AIP) – nearly invisible perturbations that are carefully tuned to confuse a target recogniser – as an aesthetic and effective anonymisation technique. Generated against a specific target recogniser, it is often hard to assess the performance when the deployed recognition system is unknown.

Oh *et al*. [3] has proposed to employ a game theoretical framework to assess the AIP performance against an unknown choice of the recogniser. Under this framework, it is possible to obtain a lower bound on the anonymisation performance independent of the choice of recogniser.

Classical obfuscation ineffective. Oh *et al*. [2] have considered the anonymisation performance of classical anonymisation techniques (figure 4) against the state of the art recogniser in §1. The authors have fine-tuned the recogniser features against a few anonymisation patterns, and utilised social media metadata (album information) to further extract cues across photos (the *Faceless Recognition* system). As a result, [2] has found that (1) face blurring and cropping out are not effective against the state of the art recogniser, and (2) even in the harshest scenario where all the faces are cropped out, and the training-testing domain shift is large (*Day* split), the recogniser performs 12× better than the no-image baseline. See [2] for extended analysis.

Figure 4. Classical anonymisation.

AIP: Adversarial Image Perturbations. AIPs are nearly invisible perturbations that are carefully crafted to confuse a target recognition system. AIPs are indeed promising as an anonymisation technique due to their aestheticity and effectiveness (figure 5 and [3]).

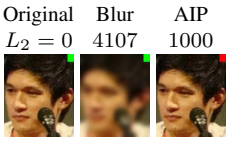

Figure 5. Anonymisation types and recognition results (green for correct, red for wrong).

However, evaluation of AIP performance is often misleading because the assumed target recogniser may not be deployed in reality, and moreover the recogniser may even adopt AIP defense techniques, an active research area.

Game theoretical framework. Game Theory provides useful tools for analysis when there exist uncertainties in the opponent players' strategies. Oh *et al*. [3] has proposed to employ a game theoretical framework to better understand the dynamics between user and recogniser (figure 6). The authors have constructed a two-player game between the user (U) and recogniser (R) striving for antagonistic goals, dis-/enabling recognition, by choosing tools for confusion/robust recognition from predefined strategy spaces. Game Theory then yields the worst-case lower bound for the obfuscation performance, independent of the choice of the deployed recognition system and defense strategies.

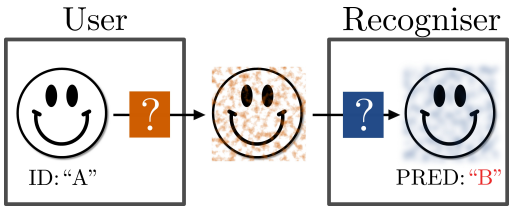

Figure 6. A game between a social media user and a recogniser over a photo. They strive for dis-/enabling recognition; they do not know which strategy is picked by the other.

Oh *et al*. [3] includes a case study of the framework. R's strategy space consists of the state of the art person recogniser in §1 as well as four efficient defense techniques applied in addition: translating (T), noising (N), blurring (B), and cropping (C), all by small amount. U's strategy space consists of the newly proposed robust AIP method `GAMAN` and four defense-resistant variants of `GAMAN` each corresponding to T, N, B, and C. Through a game theoretical analysis, [3] reports 7.3% as the worst-case upper bound on the recognition rate. It is quite low, compared to the clean image recognition rate 91.1%. See [3] for more details.

3. Conclusion

Privacy matters for social media users. [4, 1] have contributed to the understanding of identifiability in social media photos. [2] has revealed flaws in common anonymisation techniques, while [3] has suggested adversarial image perturbations as a promising alternative and proposed a game theoretical framework to account for the lack of knowledge on the deployed recognition system.

References

[1] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *ICCV*, 2015.

[2] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition; privacy implications in social media. In *ECCV*, 2016.

[3] S. J. Oh, M. Fritz, and B. Schiele. Adversarial image perturbation for privacy protection – a game theory perspective. *arXiv preprint arXiv:1703.09471*, 2017.

[4] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, 2015.