High Level Computer Vision - July 18th, 2018

Visual Privacy

Bernt Schiele - schiele@mpi-inf.mpg.de
Mario Fritz - mfritz@mpi-inf.mpg.de
Outline

Understanding privacy:
• How much private content in photos?
• Identity, places, relations, etc.?

[ICCV’15]

Controlling privacy:
• Can you “erase” content from a photo?
• .. without unpleasant effects?

[ECCV’16]

Towards a Visual Privacy Advisor:
• Understanding and predicting user’s privacy preferences

[ICCV’17]
Towards Closing the Semantic Gap

- Facilitate seamless interaction and exchange of information
- Enabling technology for applications like information retrieval, robotics and human machine collaboration
- Machines that better understand us
Towards Closing the Semantic Gap

- Facilitate seamless interaction and exchange of information
- Enabling technology for applications like information retrieval, robotics and human machine collaboration
- Machines that better understand us
Implication of Closing the Gap

- Big Data + Machine Learning is starting to close the gap
- Deep Learning: State of the art & Industrial Deployment
- Demands on Robustness, Safety, Inspection, Privacy

I need to hit the brakes!
Implication of Closing the Gap

- Big Data + Machine Learning is starting to close the gap
- Deep Learning: State of the art & Industrial Deployment
- Demands on Robustness, Safety, Inspection, **Privacy**
Privacy Implication of Personal Data Dissemination

- Shift of our lives to virtual domain
- Omnipresence of sensor and sharing online
- Social Media capture rich picture of our private lives
  - Text
  - Image
  - Video
- Pressure of industry to monetise data
- Implications of visual data dissemination on social media are not well understood
  - Not by lay persons
  - Nor by experts (no study available)
- Provide lower bound/analysis on privacy thread
- Ultimate empower and advice the user to enforce privacy policies
Machine understanding of visual content.

General object recognition

Face recognition

ILSVRC top-5 error on ImageNet

Face Recognition Performance on LFW

Source: Stanford vision lab

Source: LFW
Potential abuse.
Yet, people do and want to stay online.
Outline

Understanding privacy:
• How much private content in photos?
• Identity, places, relations, etc.?

[ICCV’15]

Controlling privacy:
• Can you “erase” content from a photo?
• .. without unpleasant effects?

[ECCV’16]

Towards a Visual Privacy Advisor:
• Understanding and predicting user’s privacy preferences  [ICCV’17]
Computer vision for human identification.

**Face recognition**
- LFW, 2007

**Pedestrian re-identification**
- PETA, 2014

**Personal photos**

**Person recognition**
- PIPA [Zhang et al. CVPR’15]
  - Flickr photos.
  - Diverse pose.
  - Diverse events & social groups.
Task: Person recognition.

- Who’s this person?
- Bounding box given.
- Closed world with \(~500\) identities.
- \(~10\) examples per identity.
[ICCV’15] Person Recognition in Personal Photo Collections.

S. J. Oh, R. Benenson, M. Fritz, B. Schiele.
Person Re-Identification in Social Networks

**Contributions**
1. New state of the art person recogniser.
2. Analysis of different cues.
3. New challenging setup.
People in Photo Albums (PIPA) dataset (Zhang CVPR’15)

- 37107 Flickr personal photo album images (Creative Commons)
- 63188 head bounding boxes
- 2356 identities
More challenging scenarios.

Event & clothing changes.

- Original [CVPR’15]
- Album
- Time
- Day
Simple and effective recogniser.

- Extract neural network features from multiple regions.
- Train identity classifiers on concatenated features (e.g. SVM).
CNN Deep Learning Architektur

Region { Face[4], Head, Upper body, Full body, Scene }


Surrogate task { default: Identity prediction, Attribute prediction }

References


[6] Benenson, Mario Fritz, Qualitative results

Linear SVM

Features

✔

✘

Split 1

Split 2

Split 3

Original [?] Album

Splits available online:

goo.gl/DKuhlY

Original split

They are in the increasing order of difficulty.


Unreasonably good performance: 33.77%, given the difficulty of the dataset.

Fixed, Album, Time, and Day splits

Original, Album, Time, and Day splits

New splits & Analysis

Body and scene cues become weaker.

Face and head cues are effective across time.

Which cues are helpful across time?

Accuracy (%)

83.88% 80.81%

83.36% 33.77%

84.88%

Evaluation protocol [1]

New challenging setup.

Head box + identity annotations.
Simple and effective recogniser.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Cues</th>
<th>Original split</th>
<th>Day split</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance level</td>
<td>-</td>
<td>0.8%</td>
<td>2.0%</td>
</tr>
<tr>
<td>[Zhang et al. CVPR’15]</td>
<td>109</td>
<td>83.1%</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>17</td>
<td>86.8%</td>
<td>46.6%</td>
</tr>
</tbody>
</table>

![Diagram showing features and test accuracy](image-url)
How many samples needed?
Example Results (day split) (goo.gl/DKuhlY)
[ICCV’15] Conclusion.

• Subjects in social media photos can be successfully identified using a simple recognition system.

• Contributions:
  2. Simple, effective method.
Outline

Understanding privacy:
• How much private content in photos?
• Identity, places, relations, etc.?

[ICCV’15]

Controlling privacy:
• Can you “erase” content from a photo?
• .. without unpleasant effects?

[ECCV’16]

Towards a Visual Privacy Advisor:
• Understanding and predicting user’s privacy preferences

[ICCV’17]
[ECCV’16] Faceless Person Recognition; Privacy Implications in Social Media.

S. J. Oh, R. Benenson, M. Fritz, B. Schiele.
Does Blurring Help?

- Blur or black-out are widespread obfuscation techniques
- How effective are they?
- Multiple images online:
  - Interplay between recognition and linking?
Does Blurring Help?

Modest reduction in recognition rate by traditional obfuscation
Up to 15 percentage points are recovered by linking

Alternatives? More effective? Visually pleasing?
ECCV’16] Conclusion.

1. Common anonymisation schemes are not effective.
2. They are visually unpleasant.
3. Ignite privacy discussion within & outside vision community.

<CVPR workshop>
Adversarial Image Perturbations

$L_2 = 0$

$L_2 = 4855$

$L_2 = 1000$

$L_2 = 153$

$L_2 = 1000$

$L_2 = 2000$

$L_2 = 3000$

Original

$L_2 = 0$

Blur

$L_2 = 2586$

GA

$L_2 = 1000$

DF\[21]\n
$L_2 = 75$

GAMAN

$L_2 = 1000$

GAMAN

$L_2 = 2000$

GAMAN

$L_2 = 3000$
“Image Gradient” Idea

• Locally linear approximation of the score function of CNN (first order Taylor approximation)

\[ S_c(I) \approx w^T I + b, \quad w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0}. \]

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps
Karen Simonyan, Andrea Vedaldi, Andrew Zisserman
Image Gradient ~ Class Specific Saliency

Gradient of Top1 Prediction
Image Gradient ~ Class Specific Saliency

Figure 3: Weakly supervised object segmentation using ConvNets (Sect. 3.2).

Left: images from the test set of ILSVRC-2013.
Left-middle: the corresponding saliency maps for the top-1 predicted class.
Right-middle: thresholded saliency maps: blue shows the areas used to compute the foreground colour model, cyan – background colour model, pixels shown in red are not used for colour model estimation.
Right: the resulting foreground segmentation masks.
Adversarial Image Perturbations (AIP).

**What:** small input noise \( t \) (human invisible), carefully crafted to confuse a target recognition system \( f \).

**How:**

\[
\max_t \mathcal{L}(f(x + t), y) \quad \text{s.t.} \quad ||t||_2 \leq \epsilon
\]

![Image of panda, adversarial noise, and resulting classification change](image)

“panda” 57.7% confidence

“nematode” 8.2% confidence

“gibbon” 99.3% confidence

[Goodfellow+, ’14]
Adversarial Image Perturbations (AIP).


AIP against deep neural nets introduced by [Szegedy+,’13].
- Perturb input directly (end-to-end architecture).
- Needed perturbation amount is sub-perceptible.

[Goodfellow+,’14]

Adversarial Image Perturbations (AIP).

+ **Aesthetic.**
+ **Effective.**

- **Defense measures:**
  active area of research.

- **Evaluation is tricky:**
  uncertainties in the deployed recogniser & possible defense measure.

![Images showing original, blur, and AIP images with corresponding $L_2$ values](image)

S. J. Oh, M. Fritz, B. Schiele.
Game Theory: modelling uncertainty.

Setup (Zero-sum game)
• Two players (User, Recogniser), antagonistic goals.
• Fix a finite set of candidate strategies for each player.
• Strategy is sampled following an unknown distribution.

+ More challenging than static, unintelligent opponent (=1 candidate strategy).
+ Principled formulation of opponent-independent guarantee.

− No guarantee for the strategies outside the fixed set.
Game Theory tutorial

- **User (U) strategy**
  Defense-specific AIPs, Diverse set of AIP algorithms, AIP generation module weights (infinite space), ...

  \[ i \in \Theta^u \]

  - AIP 1
  - AIP 2
  - AIP 3

- **Recogniser (R) strategy**
  Defense mechanism, Neural net architectures, Neural net weights (infinite space), ...

  \[ j \in \Theta^r \]

  - Def.Mech.1
  - Def.Mech.2
  - Def.Mech.3
Game Theory tutorial

- Strategy is drawn at random. The distributions are unknown to each other, and are written as

\[ \theta_u \]

\[ \theta_r \]

AIP 1
AIP 2
AIP 3

Def. Mech. 1
Def. Mech. 2
Def. Mech. 3
Game Theory tutorial

• Reward for $R_{p_{ij}}$, is the recognition rate when $U$ plays $i$ and $R$ plays $j$:

• Reward for $U$ is the mis-recognition rate (ZSG):

$$1 - p_{ij}$$
Game Theory tutorial

- R’s expected reward for randomised strategies with distributions $\theta_u$ and $\theta_r$: 

$$\sum_{i,j} \theta_i^u \theta_j^r p_{ij}$$

- Analogous for U.
Game Theory tutorial

• U’s “optimal” random strategy can be obtained by:

\[
\arg \min_{\theta^u} \max_{\theta^r} \sum_{i,j} \theta^u_i \theta^r_j p_{ij} \quad \text{s.t. } \theta^u, \theta^r \text{ are distributions.}
\]

• The optimal value, \( v \), is the value of the game.

• Interpretation: if U plays the “optimal” strategy, the recognition rate is guaranteed to be upper bounded by \( v \), independent of R’s strategy.
Case study.

- **User strategy**:
  AIP variants
  
  \{GAMAN, GAMAN\_T, GAMAN\_N, GAMAN\_B, GAMAN\_C, GAMAN\_TNBC\}.

- **Recogniser strategy**:
  Defense measures – *Translation, Noise, Blur, Crop*. [Graese et al. ICMLA’16]
  Deployed model is sampled from \{f, f\_T, f\_N, f\_B, f\_C, f\_TNBC\}.
Case study.

<table>
<thead>
<tr>
<th>User $\Theta^u$</th>
<th>Recogniser $\Theta^r$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$f$</td>
</tr>
<tr>
<td>GAMAN</td>
<td>$f_{\text{TNBC}}$</td>
</tr>
<tr>
<td>/T</td>
<td>2.5</td>
</tr>
<tr>
<td>/N</td>
<td>5.8</td>
</tr>
<tr>
<td>/B</td>
<td>0.4</td>
</tr>
<tr>
<td>/C</td>
<td>2.6</td>
</tr>
<tr>
<td>/TNBC</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Game theoretic analysis:
- Optimal U strategy: ($\text{GAMAN}_{/B}: 61\%, \text{GAMAN}_{/\text{TNBC}}: 39\%)$.
- Optimal R strategy: ($f_N: 52\%, f_N: 48\%)$.
- Guarantees < 7.4% expected rec. rate.
- (Clean image: 91.1%, No-image: 0.8%)
[ECCV'16] Conclusion.

- AIP as an effective anonymisation tool.

- Game theoretical framework to account for the lack of knowledge on the deployed recogniser.

- Robust AIP variants.
Outline

Understanding privacy:
• How much private content in photos?
• Identity, places, relations, etc.? [ICCV’15]

Controlling privacy:
• Can you “erase” content from a photo?
• .. without unpleasant effects? [ECCV’16, ArXiv’17]

Towards a Visual Privacy Advisor:
• Understanding and predicting user’s privacy preferences [ArXiv’17]
Game Theory: modelling uncertainty.

- **Antagonistic goals**: dis-/enable recognition (ZSG).
- **Deployed recogniser**: chosen from a finite set according to an *unknown* distribution.
- **Anonymisation strategy space**: example will be shown.
- **Game theory** yields an anonymisation strategy which guarantees recognition rate < V.
Outline

Understanding privacy:
• How much private content in photos?
• Identity, places, relations, etc.?

[ICCV’15]

Controlling privacy:
• Can you “erase” content from a photo?
• .. without unpleasant effects?

[ECCV’16]

Towards a Visual Privacy Advisor:
• Understanding and predicting user’s privacy preferences

[ICCV’17]
Towards a Visual Privacy Advisor

ICCV 2017

Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz
Motivation

Privacy Risk

Low

High

1
2
3
4
Motivation

Privacy Risk

Low

High
Motivation

Privacy Risk

1. Low
2. Low
3. Moderate
4. High

Insights:
Diverse users
Motivation

Privacy Risk

Low

High

Insights:
Diverse users
Privacy Leakage
Visual Privacy Advisor

Key contributions:
- Taxonomy of private information in images
- Automatic recognition of private information in images with Deep Learning
- Automatic, personalized privacy risk prediction
The Visual Privacy (VISPR) dataset

**Taxonomy** of Privacy Attributes for Visual Data

- **10 categories** - 68 Privacy attributes

**Dataset**

- 22k images, 116k labels, ~5.2 labels/image
Study 1: Understanding User Preferences

Take-aways:

1. Diverse privacy preferences
2. Few users especially sensitive to some attributes
Study 2: User’s Visual Privacy Assessment

Take-aways:

1. Inconsistent in enforcing privacy preferences
2. Especially bad for everyday images
Privacy Risk Prediction

Attributes → Annotation → User Preferences → Ground truth

Image → Attribute Predictions

Objective Task (user-independent) → Subjective Task (user-specific) → Privacy Risk Score

AP-PR → PR-CNN → Human

Slightly better at humans at the task!
Conclusions

- First taxonomy of private information in images
- First method to automatically predict broad range of privacy attributes
- Privacy risk prediction outperforms humans in several cases
  - Unawareness
  - Small objects
  - ...

CVPR’18: Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images
Privacy and Security in Machine Learning

ICLR’18: Towards **Reverse-Engineering** Black-Box Neural Networks
ArXiv’18: Understanding and Controlling User **Linkability** in Decentralized Learning

ArXiv’18: **Sequential Attacks** on Agents for Long-Term Adversarial Goals
ArXiv’18: ML-Leaks: Model and Data Independent **Membership Inference** Attacks and Defenses on Machine Learning Models
• Remarkable performance in AI - in particular image and language understanding and synthesis

• Bright and Dark Side

• Closing words (in spirit of Joseph Weizenbaum):

  Distinguish between
  • what can be done
  • what should be done
  • what decisions (choices) are reserved to humans

Joseph Weizenbaum: Computer Power and Human Reason
Thank you for your attention

Mario Fritz
Max Planck Institute for Informatics
Saarbrücken, Germany