TOWARDS A VISUAL PRIVACY ADVISOR: UNDERSTANDING AND PREDICTING RISKS IN IMAGES

Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz

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

Motivation
- Users unintentionally expose private information in images when sharing them online (e.g., Twitter, Flickr, Facebook).
- Can we extend the concept of “privacy settings” to visual content?

Abstract
We propose a Visual Privacy Advisor, an approach to enforce users’ desired privacy settings on image content. We first create a dataset of ~22k images, annotated with 68 privacy attributes. Second, we run a user study to understand privacy preferences w.r.t. to these attributes. Third, we propose models to predict user-specific privacy scores from images. Our model outperforms judgment of users, who often fail to enforce their own privacy preferences.

Dataset
- ~22k publicly available Flickr images
- Natural everyday scenes: numerous objects, often in background
- 68 Privacy Attributes: Passport, Religion, Personal Relationships, Sexual Orientation, License Plate no., etc.
- ~116k labels, with 5.22 labels per image

User Studies
Study 1: Understanding User Preferences
- Diverse preferences ⇒ Same image, different privacy risks
- Some users especially sensitive to some attributes (e.g., religion)

Study 2: Users’ Visual Privacy Assessment
- Users provide privacy risk for each attribute (x-axis). They also assess privacy risk of attributes in images (y-axis).
- Users inconsistent in enforcing their own privacy preferences.
- With everyday images (relationships, cars, landmarks), users severely underestimate privacy risk.

Privacy Attribute Prediction
- User independent multilabel classification task: Given an image, predict multiple privacy attributes.
- We compare various baseline multilabel methods for this task.
- ResNet-based model achieves an MAP of 47.45.

Personalizing Privacy Risk
Privacy Risk = \max_a (privacy rating of attribute a)
Two proposed approaches:
1. AP-PR: Uses attribute predictions and user specified privacy preferences to estimate risk
2. PR-CNN: End-to-end learning to predict user-specific risk from images. This is better at handling noisy attribute predictions

Human vs. Machine
- We compare our privacy-risk estimation approaches to users’ visual privacy assessment (from Study 2).
- Our approach achieves better Precision-Recall and L1 scores for the same images, when compared to users themselves.

Conclusion
- Users often fail to enforce their privacy preferences on images when sharing them online. Resulting implications are a major concern.
- We propose a Visual Privacy Advisor, which extends the concept of privacy settings to visual content, by providing the user a personalized privacy risk score.
- For this task, our model shows an improvement over the visual privacy assessment of users themselves.

Acknowledgements
This research was supported by the German Research Foundation (DFG CRC 1223)

References