Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images

Tribhuvanesh Orekondy
https://people.mpi-inf.mpg.de/~orekondy
Bernt Schiele
https://people.mpi-inf.mpg.de/~schiele
Mario Fritz
https://people.mpi-inf.mpg.de/~mfritz

Max Planck Institute for Informatics
Saarland Informatics Campus
Saarbrücken, Germany

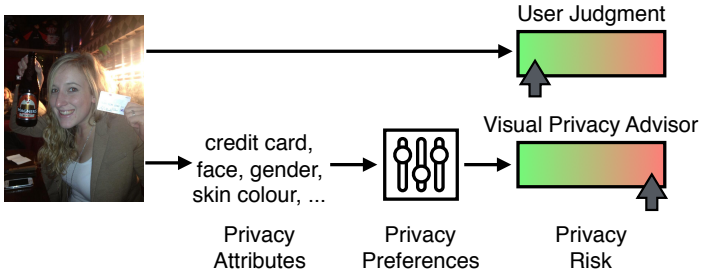

Figure 1: Users often fail to assess privacy risk while sharing images online. We propose a first *Visual Privacy Advisor*, which uses privacy attributes in an image along with the user's explicit privacy preferences to estimate the privacy risk of an image.

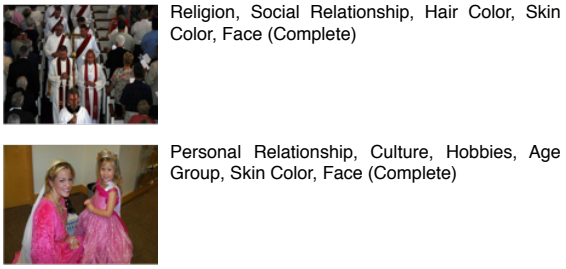

Figure 2: Example images and attribute annotations from the VISPR dataset.

1 Introduction

As more people share information on the web, a large amount of personal information becomes accessible to other users, web service providers and advertisers. To counter these problems, more and more devices (*e.g.* mobile phone) and web services (*e.g.* facebook) are equipped with mechanisms where the user can specify privacy settings to comply with his/her personal privacy preference.

While this has proven useful for explicit and textual information, we ask how this concept can generalize to visual content. While users can be asked (as we also do in our study) to specify how comfortable they are releasing a certain type of image content, the actual presence of such content is implicit in the image and not readily available for a privacy preference enforcing mechanism nor the user. In fact – as our study shows – people frequently misjudge the privacy relevant information content in an image – which leads to the failure of enforcing their own privacy preferences.

Hence, we work towards a *Visual Privacy Advisor* (Figure 1) that helps users enforce their privacy preferences and prevents leakage of private information. We approach this complex problem by first making personal information explicit by defining and predicting 68 image attributes. Based on such attribute predictions and user privacy preferences, we infer a privacy score that can be used to prevent unintentional sharing of information. Our model is trained to predict the user specific privacy risk and interestingly, it outperforms human judgment on the same images.

Our contributions are: (i) To the best of our knowledge, we are the first to formulate the problem of identifying a diverse set of personal information in images and personalizing predictions to users based on their privacy preferences (ii) We provide a sizable dataset[1] of 22k images annotated with 68 privacy attributes (iii) We conduct a user study and analyze the diversity of users' privacy preferences as well as the level to which they achieve to follow their privacy preferences on image data (iv) We propose the first model for Privacy Attribute Prediction. We also extend it to directly estimate user-specific privacy risks (v) Finally, we show that our approach to a *Visual Privacy Advisor* outperform users in following their own privacy preferences on images.

2 The Visual Privacy (VISPR) Dataset

Mobile devices and social media platforms provide privacy settings, so that users can communicate their privacy preferences on the disclosure of different types of textual information. How does this concept transfer to image data? We define and categorize personal information into 68 *privacy attributes* (*e.g.* gender, tattoo, email address, fingerprint) that an image can disclose. This will allow us to query users about their privacy preferences on the disclosure of various information types. We additionally propose a dataset (Figure 2) of 22,167 Flickr images annotated with 115,762 labels (∼5.22 attributes per image).

[1]Refer to project website: `https://tribhuvanesh.github.io/vpa/`

3 Understanding Privacy Risks

We conduct two user studies on Amazon Mechanical Turk to further understand the proposed privacy attributes. For extended analysis and discussions, we refer to [1].

3.1 Understanding Users' Privacy Preferences

In the first user study, we analyze the degree to which various users are sensitive to the privacy attributes discussed in the previous section. Hence, for each of the 305 participants of the study and the 67 privacy attributes (excluding *safe*), we ask the user how much they would find their privacy violated if they accidentally shared details of this attribute publicly online. Responses for the question are collected on a scale of 1 (Privacy not violated) to 5 (Privacy extremely violated).

We observed from this study that users show a wide variety of preferences among the attributes. Moreover, some users tend to be especially sensitive to some attributes (*e.g.* religion, sexual orientation). Hence, this supports the need for user preference-based privacy risk prediction. We refer the reader to [1] for extended analysis.

3.2 Users and Visual Privacy Judgment

In this second study, we first ask each of 50 participants to judge their personal privacy risk based on a group of 3-6 images representing an attribute (providing a visual privacy risk score) and afterwards asking the actual user's privacy preferences for the same attribute (providing a desired or explicit privacy risk score). Hence, we study how good users are at assessing their personal privacy risks based on images. We obtain responses to each of these questions on a scale of 1 (Privacy not violated) to 5 (Privacy extremely violated).

We compute for each attribute average privacy preference score and human visual scores, and visualized them as a scatter plot in Figure 3. We observe from the off-diagonal points a clear inconsistency in the users between their personal privacy preference and their judgment of privacy risk in images. Furthermore, users often severely under-estimate (below diagonal) or over-estimate (above diagonal) privacy risks of many attributes.

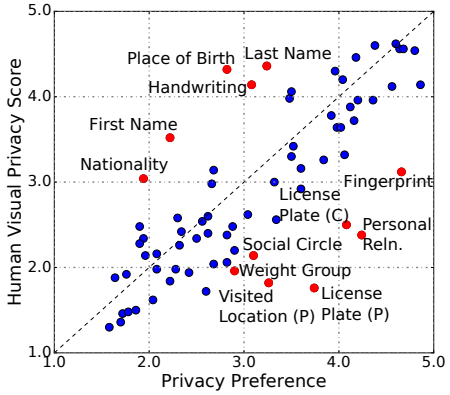

Figure 3: Users are asked to rate on a scale of 1 (Not violated) to 5 (Extremely violated) how much an attribute affects their privacy. X-axis denotes their desired privacy preference and Y-axis denotes their evaluation of risk on images. The red markers indicate privacy attributes with highly underestimated or overestimated user privacy scores.

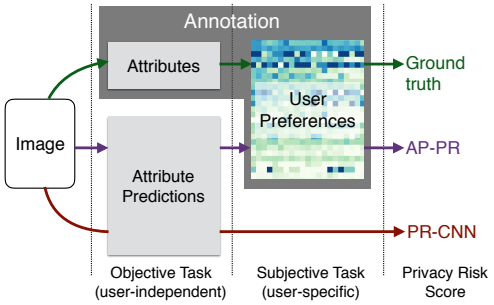

Figure 4: We learn an end-to-end model for user-specific privacy risk estimation.

4 Predicting Privacy Risks

We now make a step towards our overall goal of a *Visual Privacy Advisor*. As illustrated in Figure 4, we follow a similar paradigm *e.g.* on social networks that defines privacy risk based on both the content type and user-specific privacy settings. In our case, the content type is described by (user-independent) attributes and (user-specific) privacy preferences.

4.1 Privacy Attribute Prediction

We propose the task of *Privacy Attribute Prediction*, a multilabel classification task of predicting one or more of 68 privacy attributes based on an image. The task is challenging due to image diversity, intra-class variance, embedded text, subtle cues and high level semantics.

We approach this task by fine-tuning ImageNet pre-trained CNN models based on multi-label classification loss. We evaluate overall performance by Class-based Mean Average Precision (C-MAP), the average of the AP scores (area under Precision-Recall curves over all 68 attributes). We observe our best model obtains a C-MAP score of 47.45.

4.2 Personalizing Privacy Risk Prediction

We combine ground-truth or predicted privacy attributes (user-independent) together with privacy preferences of these attributes (user-specific) to arrive at the privacy risk score based on the following definition.

Definition 1. *Privacy Risk Score.* For some image $\boldsymbol{x}$, attributes $\boldsymbol{y} \in [0,1]^A$ and user preference $\boldsymbol{u} \in [0,5]^A$, the privacy risk score of image $\boldsymbol{x}$ containing attributes $\boldsymbol{y}$ on user $\boldsymbol{u}$ is $\max_a \boldsymbol{y}_a \boldsymbol{u}_a$

This represents the user-specific score of the most sensitive attribute, most likely to be present in an image. As a result, the privacy-risk score is comparable to the preference-score: 1 (Not Sensitive) to 5 (Extremely Sensitive). We propose two methods to predict this score from images.

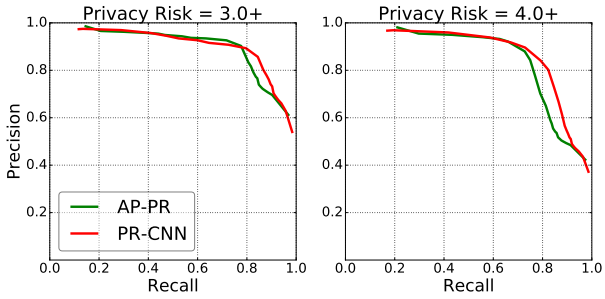

Figure 5: Performance of our approach in predicting Privacy Risks of images. Our approach performs better on high privacy-risk images.

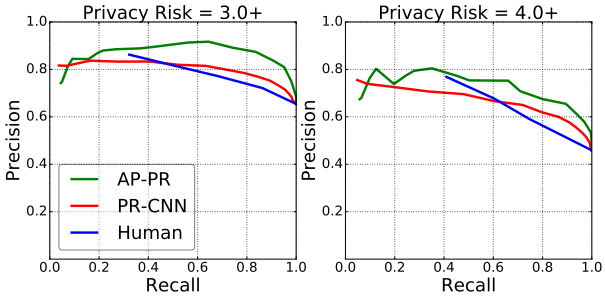

Figure 6: The Precision-Recall curves of three risk estimations are displayed – users implicitly evaluating risk from images and our two methods AP-PR and PR-CNN.

Method: Attribute Prediction-Based Privacy Risk (AP-PR) We combine the privacy attribute prediction and the profile's privacy preferences (that we can assume as provided by users at test time) to compute the privacy risk score as defined above, as shown in Figure 4.

Method: Privacy Risk CNN (PR-CNN) We propose a method that does not directly use the user's privacy preferences – but only indirectly via the ground-truth privacy risk according to Definition 1. The key observation is that AP-PR scores suffer from erroneous attribute predictions. Therefore, we extend and finetune the privacy attribute prediction network (from Section 4.1) by additional fully-connected layers to directly predict the privacy risk score.

As an evaluation metric, we calculate the Precision-Recall curves for varying thresholds of sensitivity which indicates how well our models detect images above a certain true privacy risk. Each graph in Figure 5 represents PR curves over the ground-truth thresholded to obtain a particular risk interval, such that any score above this threshold is considered private. From these results, we observe that PR-CNN performs better in predicting risk compared to using the intermediate attributes predictions. Moreover, it is better at detecting high-risk images.

4.3 Humans vs. Machine

In our user study (Section 3.2), for each attribute, users first assessed their personal privacy risk on images (providing a visual privacy risk score) and later rated their privacy preference (providing a desired privacy risk score). Now, we compare these scores with our privacy risk models AP-PR and PR-CNN on those very same images. The precision-recall-curves for the three candidates are presented in Figure 6. We observe AP-PR achieves better precision-recall for the task than PR-CNN and – remarkably – is even *consistently better than the users' image-based judgment*.

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

References

[1] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, 2017.