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

Tribhuvanesh Orekondy Bernt Schiele Mario Fritz

Max Planck Institute for Informatics
Saarland Informatics Campus
Saarbrücken, Germany
{orekondy,schiele,mfritz}@mpi-inf.mpg.de

Abstract

*With an increasing number of users sharing information online, privacy implications entailing such actions are a major concern. For explicit content, such as user profile or GPS data, devices (*e.g.* mobile phones) as well as web services (*e.g.* facebook) offer to set privacy settings in order to enforce the users' privacy preferences.*

We propose the first approach that extends this concept to image content in the spirit of a Visual Privacy Advisor. *First, we categorize personal information in images into 68 image attributes and collect a dataset, which allows us to train models that predict such information directly from images. Second, we run a user study to understand the privacy preferences of different users w.r.t. such attributes. Third, we propose models that predict user specific privacy score from images in order to enforce the users' privacy preferences. Our model is trained to predict the user specific privacy risk and even outperforms the judgment of the users, who often fail to follow their own privacy preferences on image data.*

1. Introduction

As more people obtain access to the internet, a large amount of personal information becomes accessible to *e.g.* 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.

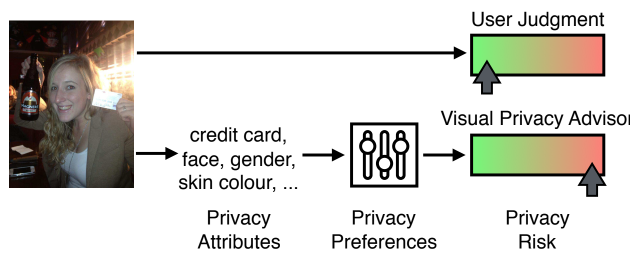

Figure 1: Users often fail to enforce their privacy preferences when sharing images online. We propose a first *Visual Privacy Advisor* to provide user-specific privacy feedback.

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 categorizing personal information into 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 main contributions in this paper are as follows: (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 models outperform users in following their own privacy preferences on images

2. Related Work

Privacy is becoming an increasing concern [48, 10], especially due to the rise of social networking websites allowing individuals to share personal information, without explaining consequences of these actions. In this section, we discuss work that highlights these concerns and explores consequences of such actions. We also discuss literature that deals with identifying private content in images and text.

Identifying Personal Information There is a comparably small body of work that aims to recognize personal information. Aura *et al*. [3] explore this in the context of electronic documents, where they propose a tool to remove user names, identifiers, organization names and other private information from text-based documents with metadata. [5, 13] study this in the context of textual email-content. Bier *et al*. [5] model this as a privacy-classification problem, whereas Geng *et al*. [13] detect four types of personal information – email addresses, telephone numbers, addresses and money. The closest related work to ours is [42], who are also motivated by unwanted disclosure and privacy violation on social media. They approach the task as classifying if an image is public or private based on features extracted from a Convolutional Neural Network and user-generated tags for the image. However, we later show that users have different notions of privacy and hence cannot be modeled as a binary classification problem. Instead, we first tackle a more principled problem of predicting the privacy-sensitive elements present in images and use these in combination with users preferences to estimate privacy risk.

Leakage and De-anonymization A problem closely related to ours is *privacy leakage*, which deals with uncovering and analyzing methods leading to disclosure of personal information, rather than detection before such incidents. [24, 22] uncover privacy leakage when websites accidentally provide user information embedded in HTTP requests when contacting third-party aggregators. As leakages can be user-intended, Yang *et al*. [47] explore this case in Android applications. Some works [32, 43] study the case where users identity, location or other details can be de-anonymized when aggregating anonymized data across multiple social networks. In contrast to these, our approach is concerned about image content and privacy preferences.

Privacy Preferences and Social Networks [26, 14, 23] study types of personal information disclosed on social networking websites. Other tasks include preserving one's privacy while using social networks [15, 52, 27] and exploring privacy settings [11, 8, 28]. However in our user study, apart from collecting and analyzing user studies on privacy preferences for images, we additionally use them to train models based on image data.

Privacy and Computer Vision Several works explore detecting individual privacy attributes such as license plates [53, 49, 6], age estimation from facial photographs [4], social relationships [45], face detection [40, 44], landmark detection [51] and occupation recognition [39]. Apart from detecting attributes, some works introduce new privacy challenges in vision such as adversarial perturbations [31, 35], privacy-preserving video capture [1, 36, 33, 37], person re-identification [2, 30], avoiding face detection [46, 16], full body re-identification [34] and privacy-sensitive lifelogging [18, 20]. In this work, we present a new challenge in computer vision designed to help users assess privacy risk before sharing images on social media that encompasses a broad range of personal information in a single study.

Datasets for Privacy Tasks Crucial to exploring privacy tasks are images revealing private details such as faces, names or opinions. However, many available datasets do not contain a significant number of such images to effectively study privacy tasks. Although some datasets [12] contain such information, they are either too small or not representative of images on social networks. The closest candidate is the PIPA dataset [50] with 37,107 Flickr images, proposed for people recognition in an unconstrained setting and does not include images covering many other privacy aspects such as license plates, political views or official identification documents. In this paper, we introduce the first dataset of real-life images capturing important privacy-relevant attributes.

3. 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 type of textual information. How does this concept transfer to image data? We need to establish a similar concept of privacy relevant information types – but now for *images*. This will allow us to query users about their privacy preferences on the disclosure of various information types, as we will do in the next section.

Therefore, we propose in this section a categorization of personal information into 68 privacy attributes such as gen-

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

der, tattoo, email address or fingerprint. We collect a dataset of 22k images that allows the study of privacy relevant attributes in images and the training of automatic recognizers.

Privacy Attributes

As motivated before, we need to categorize different types of personal content in images – akin to the privacy settings deployed in today's devices and services. Therefore, we define a list of *privacy attributes* an image can disclose.

The primary challenge here is the lack of a standard list of privacy attributes. We thus compile attributes from multiple sources. First, we consolidate relevant attributes from the guidelines for handling *Personally Identifiable Information* [29] provided in the EU Data Protection Directive 95/46/EC [9] and the US Privacy Act of 1974. Second, we add relevant attributes from the rules on prohibiting sharing personal information on various social networking websites (*e.g.*, Twitter, Reddit, Flickr). Finally, we manually examine images that are shared on these websites and identify additional attributes. As a result, we draft an initial set of 104 potential privacy attributes. As discussed in the next section, these are reduced to 68 attributes (see Table 1) after pruning.

Annotation Setup

The annotation was set up as a multi-label task to three annotators annotating independent sets of images. A web-based tool was provided to select multiple options corresponding to the 104 privacy attributes per image. Additionally, annotators could mark if they were unsure about their annotation. In case none of the provided privacy labels applied, they were instructed to label the image as *safe*, which we use as one of our privacy attributes. Images were discarded if annotators were unsure, or if the image contained a copyright watermark, was a historic photograph, contained primarily non-English text, or was of poor quality.

Data Collection and Annotation Procedure

In this section, we discuss the steps taken to obtain the final set of 22k images annotated with 68 privacy attributes.

Seed Sample We first gather 100k random images from the OpenImages dataset [21], a collection of $\sim$9 million Flickr images. Using the definition and examples of the privacy attributes, the annotators annotate 10,000 images randomly selected from the downloaded images.

Handling Imbalance Based on the label statistics from these 10,000 images, we add images to balance attributes with fewer than 100 occurrences. These additional images are added by querying relevant OpenImages labels possibly representative of insufficient privacy attributes.

Split	All	Train	Val	Test
Images	22,167	10,000	4,167	8,000
Labels	115,742	51,799	22,026	41,917
Avg Labels/Image	5.22	5.18	5.29	5.24
Max Images/Label	10,460	4,710	1,969	3,781
Min Images/Label	44	20	7	12

Table 1: Dataset Statistics

Extended Search for Rare Classes In spite of using the above strategy, 37 attributes contain under 40 images. We manually add images for these attributes by querying relevant keywords on Flickr. We do not add multiple images from the same album. For credit cards, we manually obtain 50 high-quality images from Twitter, which are the only non-Flickr images in our dataset.

Selected Attributes After annotating the dataset with the initial 104 labels, we discard 19 labels because either (i) images were difficult to obtain manually (*e.g.* iris/retinal scan, insurance details) or (ii) the set of images did not clearly represent the attribute. We additionally merge groups of attributes which capture similar concepts (*e.g.* work and home phone number). In the end, we obtain a dataset of 22,167 images, each annotated with one or more of 68 privacy attributes.

Curation To reduce labeling mistakes, we organize the dataset into batches of images with each batch corresponding to a privacy attribute. We curate attribute batches which either contain fewer than 500 images or are considered sensitive by users.

Splits We perform a random 45-20-35 split with 10,000 training, 4,167 validation and 8,000 test images. The final statistics of our dataset is presented in Table 1. The labels and its distribution in our dataset is shown in Figure 2.

4. Understanding Privacy Risks

In this section, we explore how users' personal privacy preferences relate to the attributes in Section 4.1. Furthermore, we study how good users are at enforcing their own privacy preferences on visual data when making judgments based on image data in Section 4.2.

4.1. Understanding Users' Privacy Preferences

In this section, we study the degree to which various users are sensitive to the privacy attributes discussed in Section 3.

User Study We present each user with a series of 72 questions in a randomized order. Each of these questions corresponds to either exactly one of 67 privacy attributes (excluding the safe attribute) or a control question. In each

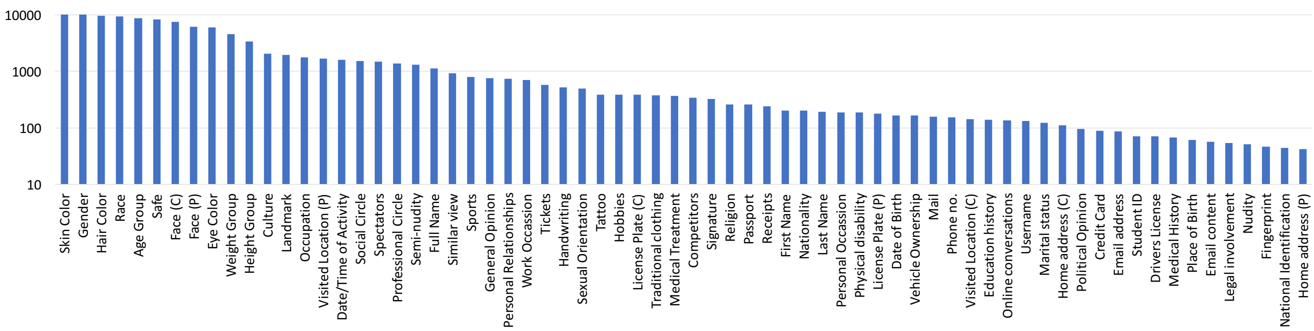

Figure 2: Label distribution in our dataset. Y-axis indicates the number of images.

question, the users are asked how much they would find their privacy violated if they accidentally shared details of a particular attribute publicly online. For instance: "How much would you find your privacy violated if you accidentally shared details on personal occasions you have attended (like a birthday party or friend's wedding)." Responses for the question are collected on a scale of 1 to 5, where: (1) Privacy is not violated (2) Privacy is slightly violated (3) Privacy is somewhat violated (4) Privacy is violated (5) Privacy is extremely violated. We treat these responses as users privacy preference for this particular privacy attribute.

Participants We collect responses of 305 unique AMT workers in this survey. Out of the 305 respondents, 59% were male, 78% were under 40 years of age with 57% from USA and 38% from India. Additionally, 75% were regular Facebook users, 80% and 44% reported to be aware of and have used Twitter and Flickr at least once.

Analysis In order to understand the diversity in users' privacy preferences, we first cluster the users based on their preferences into *user privacy profiles*. We cluster using K-means and choose K based on silhouette score [38], which considers distance between points within the cluster and additionally distance between points and their neighbouring cluster. We choose $K = 30$ as this yields the lowest silhouette score. This enables visualizing the preferences over the attributes, as seen in Figure 3, where each row represents the preferences for one of the 30 user profiles (ordered based on number of users associated with the profile). We observe from this study: (i) Users show a wide variety of preferences. This supports requiring user-specific privacy risk predictions. (ii) The *majority* (Profiles 1-4, 7-11, 13-14, 18-20 in Figure 3) display a similar order of sensitivity to the attributes (iii) A *minority* (Profiles 21-30) of users are particularly sensitive to some attributes such as their political view, sexual orientation or religion (iv) The *uniformly-sensitive users* (Profiles 5, 6, 12, 15, 17) are uniformly sensitive to all attributes even though to different degrees.

4.2. Users and Visual Privacy Judgment

In this study, we first ask participants to judge their personal privacy risk based on 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.

User Study In this study, we split the survey into two parts. In the first part, the users are shown a group of 3-6 images. Given the sensitive nature of attributes, we cannot obtain or ask users to rate their personal images and hence use images from the dataset. They are asked how comfortable they are sharing such images publicly, considering they are the subject in these images. Responses are collected on a scale of 1 to 5, where: (1) Extremely comfortable (2) Slightly comfortable (3) Somewhat comfortable (4) Not comfortable (5) Extremely uncomfortable. Each group of images represents one of the 68 privacy attributes. In most cases, the attributes occur isolated and are the most prominent visual cue in the image. We refer to these responses as *human visual privacy score*. The second part is identical to questions and the setting in the previous user-study on privacy preferences. Each question is designed to obtain the privacy preference of the user for each attribute. As before, the user rates on a scale of 1 (Not Violated) to 5 (Extremely Violated). We refer to these responses as *privacy preference score*.

Participants We split the study into two parts to prevent user fatigue. Each part contains only half of the attributes. We obtain 50 unique responses for this survey from AMT. In each of these parts, roughly: 70% of the respondents were under 40 years, 57% were male and 87% were from USA. Additionally, 80% responded that they use Facebook, 84% Twitter and 46% Flickr.

Analysis We compute for each attribute average privacy preference score and human visual scores, and visualized

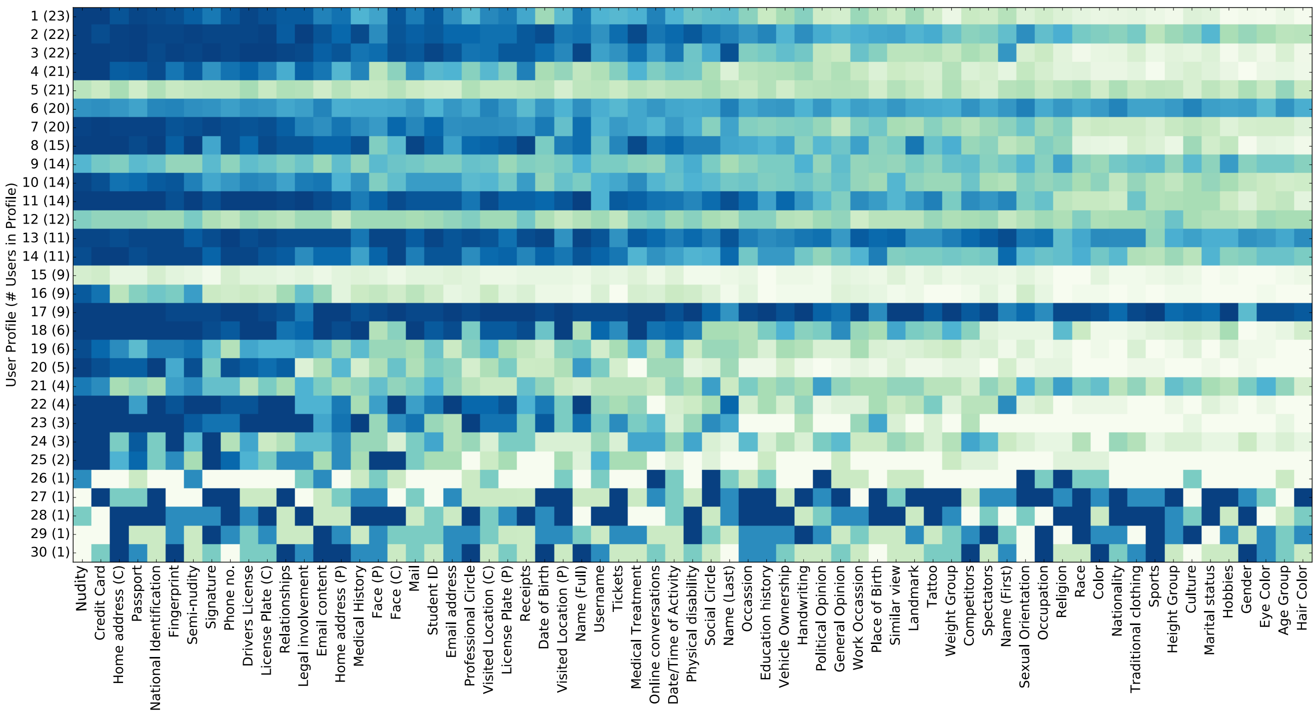

Figure 3: Privacy preferences of user profiles for the privacy attributes. Darker colors represent higher privacy-sensitivity to attributes. Each row corresponds to one of the 30 profiles and the number in brackets on the Y-axis represents the number of users mapped to the profile. Rows are ordered based on number of users linked to the profile.

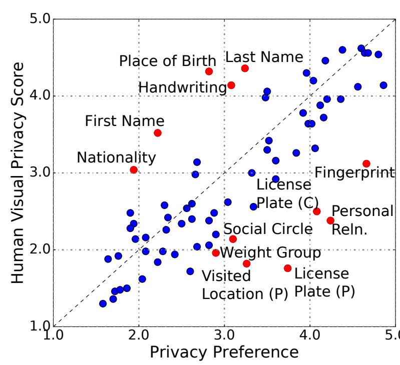

Figure 4: Users are asked to rate on a scale of 1 (Not violated) to 5 (Extremely violated) how much an attribute affects their privacy. X-axis denotes their desired privacy preference and Y-axis denotes their evaluation of risk on images. The red markers indicate privacy attributes with highly underestimated or overestimated user ratings

them as a scatter plot in Figure 4. From the results, we observe: (i) The off-diagonal data points show a clear inconsistency in the users between the required privacy preference and their judgment of privacy risk in images. (ii) For cases close to the diagonal, like credit cards, passport and national identification documents, users display consistent behaviour on images and attributes. (iii) However, when photographs are natural scenes containing people or vehicles, users underestimate (below diagonal) the privacy score, such as in the case of family photographs or cars displaying license plate numbers. We speculate this is indicative of personal photographs commonly shared online. (iv) They overestimate (above diagonal) the privacy risk of some photographs showing birth place or their name. We speculate this is because the photographs are often official documents, making users more cautious.

5. Predicting Privacy Risks

In this section, we make a step towards our overall goal of a *Visual Privacy Advisor*. As illustrated in Figure 5, 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 in the previous section. We combine these with the user-specific privacy preferences to determine if the image contains a privacy vi-

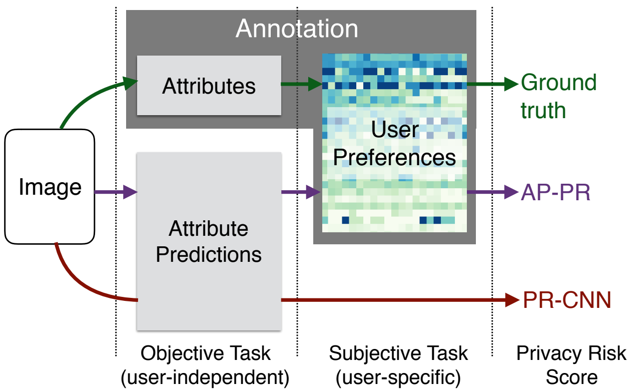

Figure 5: We learn an end-to-end model for user-specific privacy risk estimation.

olation.

We describe our model for privacy attribute prediction in Section 5.1, followed by our approaches to personalized privacy risk prediction in Section 5.2. We conclude with a comparison of human judgment of privacy risks in images against the prediction of our proposed models in Section 5.3.

5.1. Privacy Attribute Prediction

In this section, we define the *user-independent* task of predicting privacy attributes from images. Then, we present and evaluate different methods on our new VISPR dataset.

Task We propose the task of *Privacy Attribute Prediction*, which is to predict one or more of 68 privacy attributes based on an image. This can be seen as a multilabel classification problem that recognizes different type of personal information visual data and therefore has the potential to make this information explicit. Figure 1 shows multiple examples for this task. The task is challenging due to image diversity, subtle cues and high level semantics.

Metric To assess performance of methods for this task, we compute the Average Precision (AP) per class, which is the area under Precision-Recall curve for the attribute. Additionally, the overall performance of a method is given by Class-based Mean Average Precision (C-MAP), the average of the AP score across all 68 attributes.

Methods We experiment with three types of visual features extracted from CNNs – CaffeNet [19], GoogleNet [41] and ResNet-50 [17]. First, we train a linear SVM model using features from the layer preceding the last fully-connected layer of these CNNs. In a pilot study, we found that the multilabel SVM with smoothed hinge loss [25] yields better results than SVM multi-label prediction [7] and cross-entropy loss. Second, we fine-tune the CNNs initialized with pretrained ImageNet models, based on a multi-label classification loss with sigmoid activations.

Training	Features	C-MAP
SVM	CaffeNet	37.93
	GoogleNet	39.88
	Resnet-50	40.50
End-to-End	CaffeNet	42.99
	GoogleNet	43.29
	Resnet-50	47.45

Table 2: Accuracy of our methods given by Class-based Mean Average Precision, evaluated on test

Results Quantitative results of our method are shown in Table 2 and qualitative results in Figure 6 (more discussed in supplementary). We additionally present the Average Precision scores per class in Figure 7. We make the following observations: (i) The CNN performs well in attributes such as tickets, passports, medical treatment that correlated well with scenes (*e.g.* airport, hospital). It also performs well in recognizing attributes which are human–centric, such as faces, gender and age. (ii) Fine-grained differences cause confusions such as predicting student IDs for drivers licenses or differentiating between street and other signboards. (iii) We observe failure modes due to small details in the image, such as tattoos, marriage rings or a credit card in the hands of a child. (iv) Another shortcoming is not being able to recognize relationship-based attributes (*e.g.*, personal or social relationships, vehicle ownership) which requires reasoning based on interaction of multiple visual cues in an image rather than just their presence.

5.2. Personalizing Privacy Risk Prediction

In the previous section, we discussed predicting privacy attributes in images, a task independent of user privacy preferences. In this section, we investigate *user-specific* visual privacy feedback. The goal is to compute a *privacy risk score* per image, representing the risk of privacy leakage for the particular user.

Task As illustrated in Figure 5, we combine privacy attributes (user independent) together with the privacy preferences based on these attributes (user specific) to arrive at the privacy risk score. As we allow the users to give scores for each attributes based on their privacy preferences, we define the following *privacy risk score*.

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 re-

Figure 6: Qualitative Results of our Privacy Attribute Prediction method

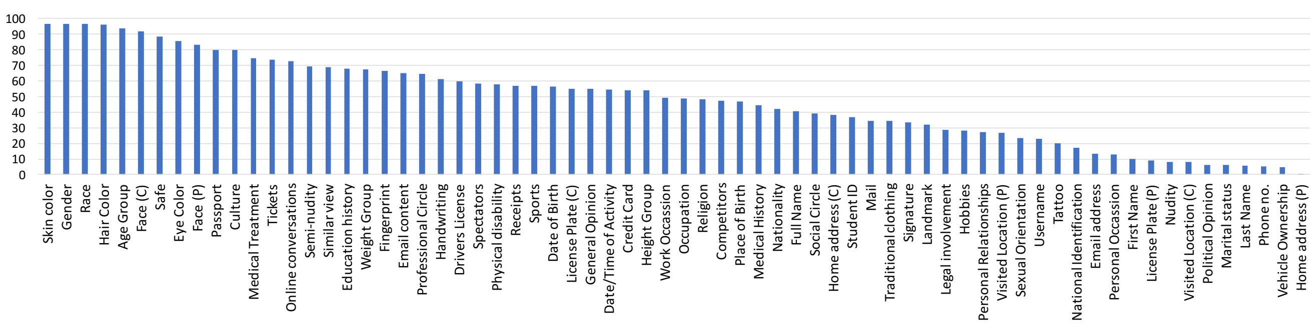

Figure 7: Average Precision (AP) Scores for the privacy attributes from our method

sult, the privacy-risk score is comparable to the preference-score: 1 (Not Sensitive) to 5 (Extremely Sensitive). As illustrated in Figure 5, we compute the ground-truth privacy risk score based on ground-truth attribute annotation for an image (represented as a k-hot vector $\boldsymbol{y} \in \{0,1\}^A$) and privacy preferences of users.

Method: Attribute Prediction-Based Privacy Risk (AP-PR) Our first method performs Attributed-Based Privacy Risk (*AP-PR*) prediction. As illustrated in Figure 5, 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.

Method: Privacy Risk CNN (PR-CNN) We propose a Privacy Risk CNN (*PR-CNN*) that does not directly use the user profile's privacy preferences – but only indirectly via the ground-truth. The key observation is that AP-PR scores suffer from erroneous attribute predictions (see Figure 7). Therefore, we extend the the privacy attribute prediction network by additional fully-connected layers to directly predict the privacy risk score. A parameter search yielded best results using additional two fully-connected hidden layers of 128 neurons, each followed by sigmoid activations. We finetune this network from our Googlenet Privacy Attribute Prediction network for 30 user profiles described in Section 4 and a Euclidean loss.

	L1-Error	MAP			
		1+	2+	3+	4+
AP-PR	0.656	**94.94**	**94.27**	87.97	77.89
PR-CNN	**0.637**	94.35	93.65	**88.14**	**78.38**

Table 3: Evaluation of Personalized Privacy Risk

Evaluation We use two metrics for evaluation. First, the $L1$ error averaged over all images and profiles; it represents the mean absolute difference between the ratings. Secondly, 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. By calculating the area under the Precision-Recall curves over all user profiles, we additionally report the Mean Average Precision (MAP).

In our experiments, we use the previously introduced user-profiles instead of individual users in order to cater to

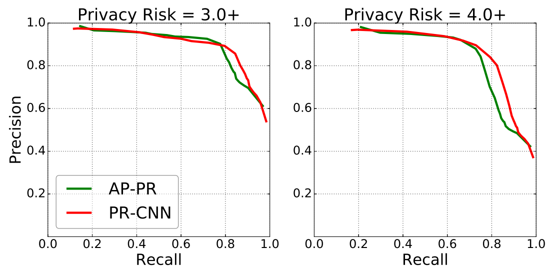

Figure 8: Performance of our approach in predicting Privacy Risks of images. Our approach performs better on high privacy-risk images.

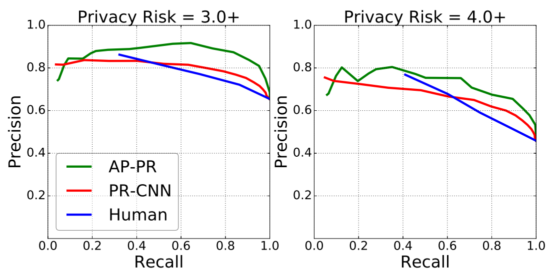

Figure 9: The Precision-Recall curves of three risk estimations are displayed – users implicitly evaluating risk from images and our two methods AP-PR and PR-CNN.

all the diverse privacy preferences equally that we have seen in the previous section. We assign a privacy risk score of 0.5 for the *safe* attribute for all profiles.

The evaluation of our approach on these metrics is presented in Table 3. Each graph in Figure 8 represents PR curves over the ground-truth thresholded to obtain a particular risk interval, such that any score above this threshold is considered private. This allows us to estimate performance of methods at various levels of sensitivity. We then obtain the PR-curves for each sensitivity interval by thresholding scores estimated by AP-PR and PR-CNN.

From these results, we observe: (i) PR-CNN performs better in predicting risk compared to using the intermediate attributes predictions. Notably, the prediction is on average less than one step on the scale from 1 to 5 away from the true privacy risk. (ii) Moreover, it is better at detecting high-risk images, as shown in Figure 8. In particular, we notice better recall for high-risk images. We discuss profile-specific PR curves in the supplementary material.

5.3. Humans vs. Machine

In Section 4, we have shown inconsistency in users' privacy preferences and their assessment of privacy risks in images. In this section, we compare our proposed approach for evaluating privacy risk against human judgments.

In our second user study (subsection 4.2), 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). We have computed scores with our privacy risk models AP-PR and PR-CNN on those very same images.

As a result, for each image, we have (a) users' privacy preference (b) users' privacy risk judgment from images (c) our AP-PR privacy risk score from images (d) our PR-CNN privacy risk score from images. All these scores are on a scale of 1 (Not Sensitive) to 5 (Extremely Sensitive). Using the users desired preference as the ground-truth, we now ask: *who is better at reproducing the user's desired privacy preference on images?* As from the previous section, we use precision-recall and L_1-error as metrics to compare the desired preference score (a) and predicted privacy risk score for evaluation (b, c, d).

The precision-recall-curves for the three candidates are presented in Figure 9. Evaluation using the L_1-error is discussed in the supplementary material. We observe: (i) AP-PR achieves better precision-recall for the task than PR-CNN and – remarkably – is even *consistently better than the users' image-based judgment*. (ii) On average, the PR-CNN estimates privacy risks (L_1 error = 1.03) slightly better than the user's image-based judgment (L_1 error = 1.1) and AP-PR (L_1 error = 1.27).

6. Conclusion

We have extended the concept of privacy settings to visual content and have presented work towards a *Visual Privacy Advisor* that can provide feedback to the users based on their privacy preferences. The significance of this research direction is highlighted by our user study which shows users often fail to enforce their own privacy preferences when judging image content. Our survey also captures typical privacy preference profiles that show a surprising level of diversity. Our new VISPR dataset allowed us to train visual models that recognize privacy attributes, predict privacy risk scores and detect images that conflict with user's privacy. In particular, a final comparison of human vs. machine prediction of privacy risks on images, shows an improvement by our model over human judgment. This highlights the feasibility and future opportunities of the overarching goal – a *Visual Privacy Advisor*.

Acknowledgement

This research was supported by the German Research Foundation (DFG CRC 1223). We thank Paarijaat Aditya, Philipp Müller and Julian Steil for advice on the user study. We also thank Dr. Mykhaylo Andriluka and Seong Joon Oh for valuable feedback on the paper.

References

[1] P. Aditya, R. Sen, P. Druschel, S. Joon Oh, R. Benenson, M. Fritz, B. Schiele, B. Bhattacharjee, and T. T. Wu. I-pic: A platform for privacy-compliant image capture. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 235–248. ACM, 2016. 2

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015. 2

[3] T. Aura, T. A. Kuhn, and M. Roe. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 41–50. ACM, 2006. 2

[4] C. Bauckhage, A. Jahanbekam, and C. Thurau. Age recognition in the wild. In *ICPR*, 2010. 2

[5] C. Bier and J. Prior. Detection and labeling of personal identifiable information in e-mails. In *IFIP International Information Security Conference*, pages 351–358. Springer, 2014. 2

[6] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen. Automatic license plate recognition. *IEEE Trans. Intelligent Transportation Systems*, 5:42–53, 2004. 2

[7] K. Crammer and Y. Singer. A family of additive online algorithms for category ranking. *Journal of Machine Learning Research*, 3(Feb):1025–1058, 2003. 6

[8] G. Danezis. Inferring privacy policies for social networking services. In *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*, pages 5–10. ACM, 2009. 2

[9] E. Directive. 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the EC*, 23(6), 1995. 3

[10] C. Dwyer, S. Hiltz, and K. Passerini. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. *AMCIS 2007 proceedings*, page 339, 2007. 2

[11] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web*, pages 351–360. ACM, 2010. 2

[12] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Proc. CVPR*, 2008. 2

[13] L. Geng, L. Korba, X. Wang, Y. Wang, H. Liu, and Y. You. Using data mining methods to predict personally identifiable information in emails. In *International Conference on Advanced Data Mining and Applications*, pages 272–281. Springer, 2008. 2

[14] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80. ACM, 2005. 2

[15] S. Guha, K. Tang, and P. Francis. Noyb: Privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, pages 49–54. ACM, 2008. 2

[16] A. Harvey. Cv dazzle: Camouflage from computer vision. *Technical report*, 2012. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[18] R. Hoyle, R. Templeman, D. Anthony, D. Crandall, and A. Kapadia. Sensitive lifelogs: A privacy analysis of photos from wearable cameras. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1645–1648, New York, NY, USA, 2015. ACM. 2

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 6

[20] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia. Enhancing lifelogging privacy by detecting screens. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2016. 2

[21] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 3

[22] B. Krishnamurthy, K. Naryshkin, and C. E. Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web*, volume 2, pages 1–10, 2011. 2

[23] B. Krishnamurthy and C. E. Wills. Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, pages 37–42. ACM, 2008. 2

[24] B. Krishnamurthy and C. E. Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 7–12. ACM, 2009. 2

[25] M. Lapin, M. Hein, and B. Schiele. Loss functions for top-k error: Analysis and insights. In *CVPR*, 2016. 6

[26] A. Lenhart and M. Madden. Teens, privacy and online social networks: How teens manage their online identities and personal information in the age of myspace. 2007. 2

[27] M. Li, N. Cao, S. Yu, and W. Lou. Findu: Privacy-preserving personal profile matching in mobile social networks. In *INFOCOM, 2011 Proceedings IEEE*, pages 2435–2443. IEEE, 2011. 2

[28] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 61–70. ACM, 2011. 2

[29] E. McCallister. *Guide to protecting the confidentiality of personally identifiable information*. Diane Publishing, 2010. 3

[30] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person reidentification. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 1325–1334, 2016. 2

[31] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *arXiv preprint arXiv:1610.08401*, 2016. 2

[32] A. Narayanan and V. Shmatikov. De-anonymizing social networks. *CoRR*, abs/0903.3276, 2009. 2

[33] C. Neustaedter, S. Greenberg, and M. Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(1):1–36, 2006. 2

[34] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016. 2

[35] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016. 2

[36] F. Pittaluga and S. J. Koppal. Privacy preserving optics for miniature vision sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 314–324, 2015. 2

[37] N. Raval, A. Srivastava, K. Lebeck, L. Cox, and A. Machanavajjhala. Markit: Privacy markers for protecting visual secrets. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1289–1295. ACM, 2014. 2

[38] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 4

[39] M. Shao, L. Li, and Y. Fu. What do you do? occupation recognition in a photo via social context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3638, 2013. 2

[40] X. Sun, P. Wu, and S. C. H. Hoi. Face detection using deep learning: An improved faster rcnn approach. *CoRR*, abs/1701.08289, 2017. 2

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 6

[42] A. Tonge and C. Caragea. Privacy prediction of images shared on social media sites using deep features. *arXiv preprint arXiv:1510.08583*, 2015. 2

[43] M. H. Veiga and C. Eickhoff. Privacy leakage through innocent content sharing in online social networks. *arXiv preprint arXiv:1607.02714*, 2016. 2

[44] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2001. 2

[45] G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, 2010. 2

[46] M. J. Wilber, V. Shmatikov, and S. Belongie. Can we still avoid automatic face detection? In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016. 2

[47] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang. Appintent: analyzing sensitive data transmission in android for privacy leakage detection. In *ACM Conference on Computer and Communications Security*, 2013. 2

[48] A. L. Young and A. Quan-Haase. Information revelation and internet privacy concerns on social network sites: a case study of facebook. In *Proceedings of the fourth international conference on Communities and technologies*, pages 265–274. ACM, 2009. 2

[49] H. Zhang, W. Jia, X. He, and Q. Wu. Learning-based license plate detection using global and local features. In *ICPR*, 2006. 2

[50] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[51] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[52] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 506–515. IEEE, 2008. 2

[53] W. Zhou, H. Li, Y. Lu, and Q. Tian. Principal visual word discovery for automatic license plate detection. *IEEE Trans. Image Processing*, 21:4269–4279, 2012. 2