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

Supplementary Material

Tribhuvanesh Orekondy Bernt Schiele Mario Fritz

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

A. Privacy Attributes and Examples

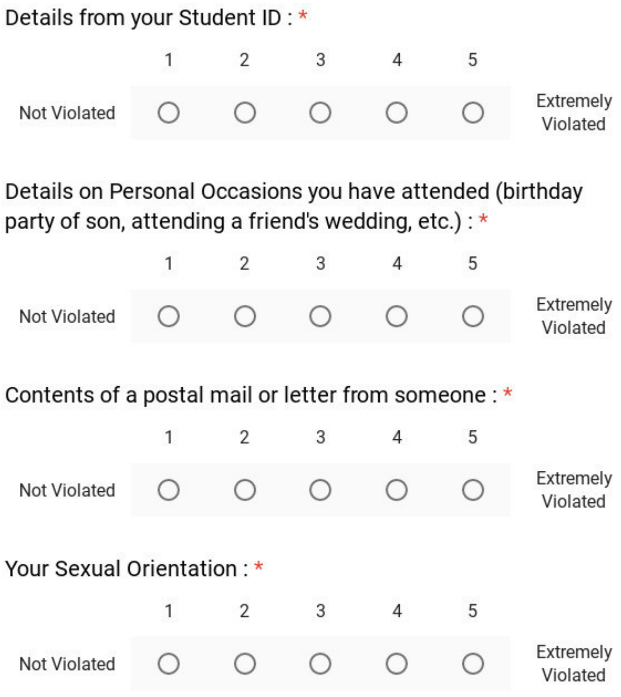
A complete list of privacy attributes with descriptions and an example image is given in Table 1. We consider all these cases when viewing the image in its original high-resolution form. We use these definitions to any subject in the image – either in the foreground or background. Using these definitions, attributes can be typically inferred from an image in multiple ways: (a) *Direct*: it is explicitly mentioned, such as in a form or document (*e.g.* gender on an identity card) (b) *Visual*: based on visual cues (*e.g.* gender from clothing or facial features) (c) *Reasoning*: it is inferred by some additional reasoning (*e.g.* relationships based on age differences between multiple people). Dataset is available on the project website: `https://tribhuvanesh.github.io/vpa/`.

B. Additional Details on User Study

In this section, we provide additional details on the user study discussed in Section 4.

B.1. Understanding Users' Privacy Preferences

The task in this user study is to obtain user preferences over the 67 privacy attributes (excludes the attribute *safe*). The questionnaire instructs the user on a fictitious website (similar to Flickr or Twitter), where content posted is by default visible to everyone else on the platform. By unintentionally posting information about a particular attribute, the user exposes private information comprising his/her anonymity. Each question is a verbal description of one of the attributes (Figure 1). We collect responses on a scale of 1-5 of how much the user finds his/her privacy violated as a consequence of this action.



Details from your Student ID : *

	1	2	3	4	5	
Not Violated	○	○	○	○	○	Extremely Violated

Details on Personal Occasions you have attended (birthday party of son, attending a friend's wedding, etc.) : *

	1	2	3	4	5	
Not Violated	○	○	○	○	○	Extremely Violated

Contents of a postal mail or letter from someone : *

	1	2	3	4	5	
Not Violated	○	○	○	○	○	Extremely Violated

Your Sexual Orientation : *

	1	2	3	4	5	
Not Violated	○	○	○	○	○	Extremely Violated

Figure 1: Questions from user study to understand privacy preferences

Instructions provided to the Users

In this academic survey we want to understand how sensitive you are to certain details of your personal or private life. For instance, are you more comfortable sharing your full name, gender or details on your personal relationships?

We refer to these details of your personal or private life

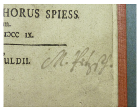
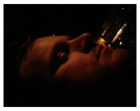
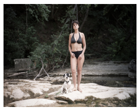
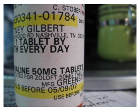
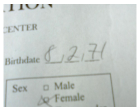
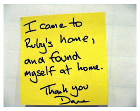
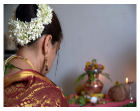
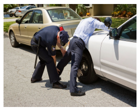
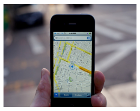
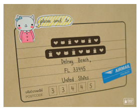
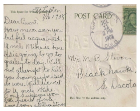
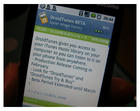
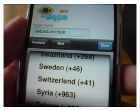
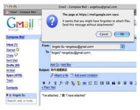
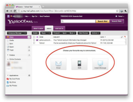
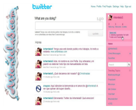
Group	Attribute	Description	Examples
Personal Description	Gender	Subject's gender is clearly visible using one or more gender-specific discriminative visual cues such as more than 50% body being visible, clothing, facial/head hair or colored nails.	
	Eye Color	If eyes are visible and can be categorized as one of: brown, hazel, blue or green.	
	Hair Color	Subject's head hair color is visible	
	Fingerprint	Fingerprint is visible through either a close-up shot of one's finger or an imprint on some surface.	
	Signature	Complete signature is visible in an image, such as in a form or document	
	Face (Complete)	A face is completely visible. Also includes photographs of faces on identity cards, documents or billboards.	
	Face (Partial)	Less than 70% of the face is visible or there is occlusion, such as when the subject is wearing sun-glasses.	
	Tattoo	Subject displays either a tattoo or body paint.	
	Nudity (Partial)	Subject appears in undergarments	
	Nudity (Complete)	Human subject appears without clothing	
	Race	Any subject in the photograph can be categorized into one of Caucasian, Asian or Negroid.	
	(Skin) Color	One's skin color can be categorized into one of White, Brown or Black.	
	Traditional Clothing	Subject appears in clothing which is indicative of a particular region or country *e.g.* dirndl, sari.	

Table 1: List of Privacy Attributes including their definitions and examples

Group	Attribute	Description	Examples
	Full Name	A recognizable full name which appears in the context of a form, document or a badge. Also includes if the name can be inferred from a signature.	Jeff Clavier SoftTech VC Business Angel Startup Advisor Hacked in C/C++, CTO/VP Engineering General Partner at Reuters Venture Capital Started investing in Web 2.0 3 years ago 20+ investments, 3 exits More: softtechvc.com, blog.softtechvc.com
	Name (First)	Only if the first name is visible on a form, document, badge or clothing.	
	Name (Last)	Only if the last name is visible on a form, document, badge or clothing.	 C. STOBER 60341-01784 EY GILBERT NASHVILLE, TN TABLET BY EVERY DAY LINE 50MG TABLET MFG GREEN BEFORE 05/08/0
	Place of Birth	Place of Birth is explicitly mentioned, such as in a form or in an identification document.	
	Date of Birth	Date of Birth is explicitly mentioned in writing. Includes year, month or the day of birth.	
	Nationality	A passport indicating country is clearly visible. Includes the case if a subject appears holding a country's flag or wearing a uniform bearing the flag (such as a soldier or an international athlete).	
	Handwriting	Hand-written text on any surface.	 I came to Ruby's home, and found myself at home. Thank you Dana
	Marital status	A subject is wearing an engagement ring. Includes wedding photographs taken of the bride and groom.	
Documents	National Identification	Documents such as a Green Card or a European national identity card, not including passports.	
	Credit Card	Either the front or back of a credit card. Includes cases when the card is partially visible *e.g.* in someone's hand or in a shredded form	
	Passport	A photograph of any page in the passport or its front cover.	
	Drivers License	Either front or back of a drivers license or a driving permit.	
	Student ID	Front or back of a student identity card, with at least the name of a school, college or university clearly readable.	

Group	Attribute	Description	Examples
	Mail	Contents of a mail or the envelope.	
	Receipts	Purchase receipts indicating a financial transaction with an amount clearly visible, *e.g.* a restaurant receipt.	
	Tickets	A travel, movie or concert ticket which specifies travel location or an event.	
Health	Physical disability	Subject appears with a permanent physical disability *e.g.* an amputee or a person in a wheelchair.	
	Medical Treatment	Subject appears either with an injury or indicates hospital admittance.	
	Medical History	Photographs of medicine or medical prescriptions.	
Employment	Occupation	Subject appears in a distinguishable occupation-specific uniform *e.g.* doctor, policemen, construction worker.	
	Work Occasion	Subject is photographed while giving a talk, presentation, attending a work-related or broad-casting event. Includes photographs of people in formal attire in an office.	
Personal Life	Religion	Subject appears associated with a distinguishable religious symbol, religion-specific clothing or at a religious location.	
	Sexual Orientation	Two subjects are photographed in an intimate setting	
	Culture	Subjects appear celebrating a traditional festival or attending an art or culture related activity *e.g.* concert, play.	
	Hobbies	A non-professional related activity of a subject is visible *e.g.* playing a musical instrument, taking photographs.	
	Sports	Subject appears taking part in an indoor or outdoor sports activity	

Group	Attribute	Description	Examples
	Education history	Photographs contains cues indicating subject's education history, such as at a graduation ceremony, clothing indicating university or an academic or school certificate	
	Legal involvement	Photographs indicating subject's involvement with law-related activities *e.g.* someone being arrested, in a court hearing.	
	Personal Occasion	Photographs of people celebrating a personal occasion with friends or family members *e.g.* wedding, birthday.	
	General Opinion	Subject appears associated with a placard or clothing indicating opinion on general issues *e.g.* wars, taxes, LGBT rights.	
	Political Opinion	Subject appears with either clothing, placard or in a crowd at a political rally.	
Relationships	Personal Relationships	Photographs of people in a visually-identifiable personal relationship *e.g.* mother-son, husband-wife.	
	Social Circle	Subjects of the same age-group photographed in a casual setting *e.g.* friends at a party, walking together on a street.	
	Professional Circle	A group of people who share an occupation (*e.g.* a group of policemen) or who are dressed for a professional event (*e.g.* a conference or meeting).	
	Competitors	A group of people taking part in team sports. Also includes the case when subjects belong to the same team.	
	Spectators	A group of people spectating an event such as a concert or play.	
	Similar view	A group of people at a rally or a protest who share opinions on a general issue. Only includes the case when placards or clothing denoting a cause or rallying for a political party is visible.	
Whereabouts	Visited Landmark	Photograph contains text indicating a business' name, street sign or a well-known landmark.	

Group	Attribute	Description	Examples
	Visited Location (Complete)	Text indicating a *complete* address (*e.g*. restaurant receipt with the address of the restaurant) or a screen-shot of GPS-based location.	
	Visited Location (Partial)	Text which partially indicates the subject's location, such as street name, city or country where the photograph was taken.	
	Home address (Complete)	Photograph containing a complete non-commercial postal address.	
	Home address (Partial)	Photograph containing a partial non-commercial postal address.	
	Date/Time of Activity	Photograph contains information of date and/or time of subject's location or activity such as a time-stamp watermark in an image, or a clock in the photograph.	
	Phone no.	A phone number that is visible in the photograph (either personal or commercial).	
Internet Activity	Username	A screen shot of a website which mentions any username or internet handles.	
	Email address	Any complete valid email-address that appears in a photograph or a screen-shot.	
	Email content	Screenshots of emails including the subject of the email, or parts of the email body content.	
	Online conversations	Screenshots of online conversations, posts, tweets or internet activity by any user.	
Automobile	Vehicle Ownership	Photograph of a person riding a motor vehicle.	
	License Plate (Complete)	A clearly visible license plate or registration number of any motor vehicle.	
	License Plate (Partial)	A partial license plate or registration number of any motor vehicle	

as "Personally Identifiable Information" (PII).

PII is information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual. Such information could be one or more of your: Full Name, Home Address, Political Opinion, etc.

Following this description are a list of PIIs. For each of these PIIs, consider the following situation: On an online public platform, you create an anonymous account. On this platform, once you post something, you cannot delete it. Only the moderators can delete this post. However, they can be extremely slow and unresponsive. One day, you unintentionally shared/posted this PII about yourself. Immediately, you realize that you cannot delete this post.

On a scale of 1-5, please rate how much you feel your privacy is violated by this action, where:
1 - I feel my privacy is not violated. So, I wouldn't care.
2 - I feel my privacy is slightly violated. However, it's not worth taking any action.
3 - I feel my privacy is somewhat violated. I will message the moderator. In case there's no response, I will give up.
4 - I feel my privacy is violated. I will inform the moderator and follow up for a few days. In case there's no response after that, I will give up.
5 - I feel my privacy is extremely violated. I will not give up until this post is deleted.

B.2. Users and Visual Privacy Judgment

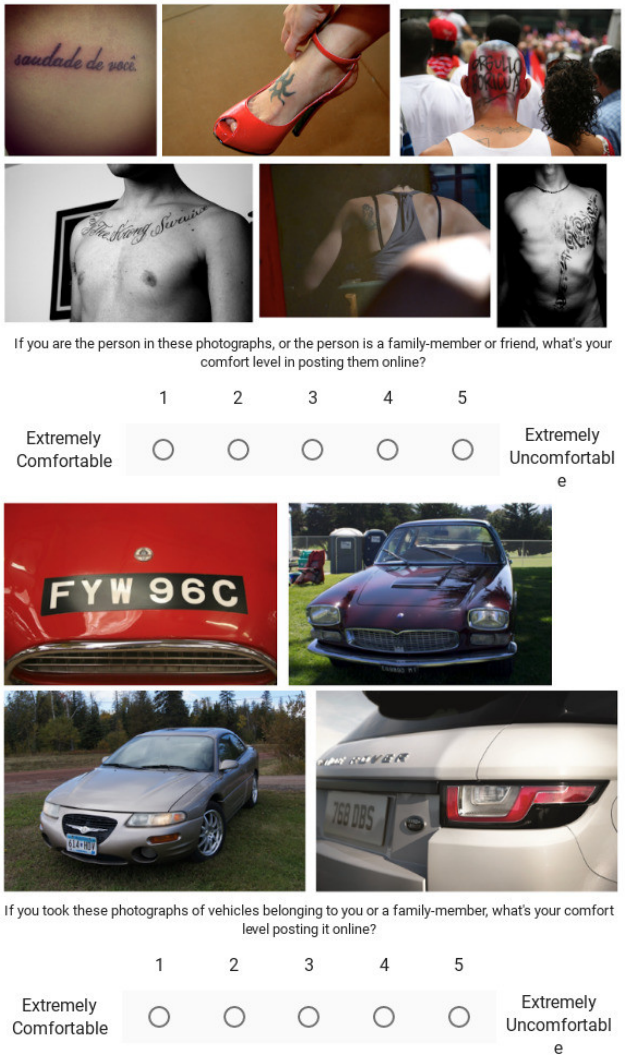
In order to understand how good are users at identifying privacy risks from images, we conduct this user study in two parts. In the first part, we instruct users on a fictitious photo-sharing website, where images shared are publicly available. For each of the 68 privacy attributes, we present a question on a group of images from the dataset representing this attribute (Figure 2). The user responds how comfortable he/she is posting such images on the website. The exact instructions for this part is provided below.

In the second part, we obtain user preferences over the attributes following the exact instructions in the previous section.

Instructions provided to the Users

In this academic survey we want to understand your comfort level sharing things on the internet.

Following this description are groups of images. For each of these groups of images, consider the following situation: On an online public platform, you create an account. On this platform, you are allowed to post photographs, which anyone can view. Moreover, you can also interact with other users who shared their photographs and can comment on or like them.



If you are the person in these photographs, or the person is a family-member or friend, what's your comfort level in posting them online?

1 2 3 4 5

Extremely Comfortable ○ ○ ○ ○ ○ Extremely Uncomfortabl e

If you took these photographs of vehicles belonging to you or a family-member, what's your comfort level posting it online?

1 2 3 4 5

Extremely Comfortable ○ ○ ○ ○ ○ Extremely Uncomfortabl e

Figure 2: Questions from the user study to evaluate user privacy judgment

Important: For each of the below groups of images, picture yourself as either being the subject in the photograph, or the one who took the photograph of a family-member.

On a scale of 1-5, rate how comfortable you are sharing such photographs, where:
1 - You are extremely comfortable sharing such photographs
2 - You are slightly comfortable sharing such photographs
3 - You are somewhat comfortable sharing such photographs
4 - You are not comfortable sharing such photographs
5 - You are extremely uncomfortable sharing such photographs

C. Additional Qualitative Examples for Privacy Attribute Prediction

In Section 5.1 we discussed our approach to *Privacy Attribute Prediction* – a user-independent method of predicting multiple privacy attributes given an image. In this section, in addition to Figure 6, we present additional qualitative examples in Figure 3. Each row represents images of a particular privacy attribute. The True Positives column indicate the case when this attribute is in both the ground-truth and predicted set of privacy attributes. The False Positives column indicate images when the attribute is incorrectly predicted. The False Negatives column indicate images when the attribute is in ground-truth, but is not predicted.

We observe our method associates privacy attributes to distinctive visual cues such as clothing (for occupation and ethnic clothing), exposed skin (for tattoos, nudity), metallic objects with wheels (for physical disability, license plates) and text (for names, drivers license, username, handwriting). As a result, apart from correct predictions, we find that this also leads to incorrectly predicting attributes (*e.g.* predicting card-shaped identification documents as drivers licenses, cars for license plates) or failing to recognize attributes in a different context (*e.g.* handwriting on a wall instead of documents, new types of drivers licenses). We also observe our approach underperform in differentiating between full, first and last names, or usernames and email addresses (which requires text-based reasoning), identifying relationships and sexual orientation (which requires interpreting interaction between multiple people) and differentiating occupations, religion and ethnic clothing (which requires fine-grained recognition).

D. Additional Results for Personalized Privacy Prediction

D.1. Qualitative Results

In this section, we discuss additional results for Section 5.2: Personalizing Privacy Risk Prediction.

Figure 4 presents qualitative results for our approach to user-specific *Personalized Privacy Risk Prediction* discussed in Section 5.2. To visualize the qualitative results over all 30 user profiles simultaneously, we present a scatter plot of ground-truth vs. predicted scores for each image. Each point in the scatter plot represents one user-profile. In these plots, points closer to the diagonal (dotted line) indicate lower errors. Points above the diagonal indicate risk over-estimation and under the diagonal indicate risk under-estimation.

We observe from the qualitative results and w.r.t each row in Figure 4: (i) (First row) presents examples with correct high confidence attribute predictions according to the posterior probability. Here, both AP-PR and PR-CNN perform equally well. (ii) (Second row) presents examples where attribute predictions are noisy. In these, PR-CNN outperforms AP-PR. (iii) (Third row) Both AP-PR and PR-CNN are challenged by difficult images (low contrast, unnatural angles, low lighting, occlusion). However, we see that PR-CNN often performs slightly better than AP-PR in these cases. (iv) (Fourth row) presents examples where AP-PR with correct attribute predictions performs better than PR-CNN.

D.2. Precision-Recall Curves for User Profiles

Section 5.2 discussed Precision-Recall curves evaluated over all profiles. These were obtained by treating the privacy risk-prediction as a binary classification problem, where images above a certain risk score (3+ and 4+ previously) is considered private per user profile.

In Figure 5, we present the Precision-Recall curves evaluated over groups of profiles and additional risk thresholds. To generate the curves in these figures, we first create four groups of profiles, with an equal number of profiles in each group. We refer to these groups as quartiles Q1-Q4. We then obtain the Precision-Recall curves for each of these quartiles.

We observe that PR-CNN displays better performance for high-risk images over *all* quartiles of the 30 user profiles and hence contributing to an overall better performance.

Additionally, we observe a similar pattern with the L_1-error metric (the absolute difference in scores), where PR-CNN (error = 0.67) incurs lower error in scores for private images compared to AP-PR (error = 0.84). However, AP-PR (error = 0.34) performs better for safe images in comparison to PR-CNN (error = 0.58).

E. Additional Results for Humans vs. Machine

In Section 5.3, we discussed the performance of our Privacy Risk Evaluation Methods when compared to the users themselves. The performance evaluation was primarily with Precision-Recall curves.

In this section, we discuss performance when evaluated using L_1 as a distance metric between the ground-truth privacy scores (user's specified preferences) and the privacy risk estimation using three approaches (user's visual risk assessment and our two proposed approaches – AP-PR and PR-CNN). The L_1 distance here measures the absolute difference in risk score (where risk scores are between 1–5). Figure 6 presents these errors per attribute.

We observe from these results: (i) On average (horizontal lines), 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) (ii) Users often misjudge the risk (right end of figure) from natural-looking images such as cars with

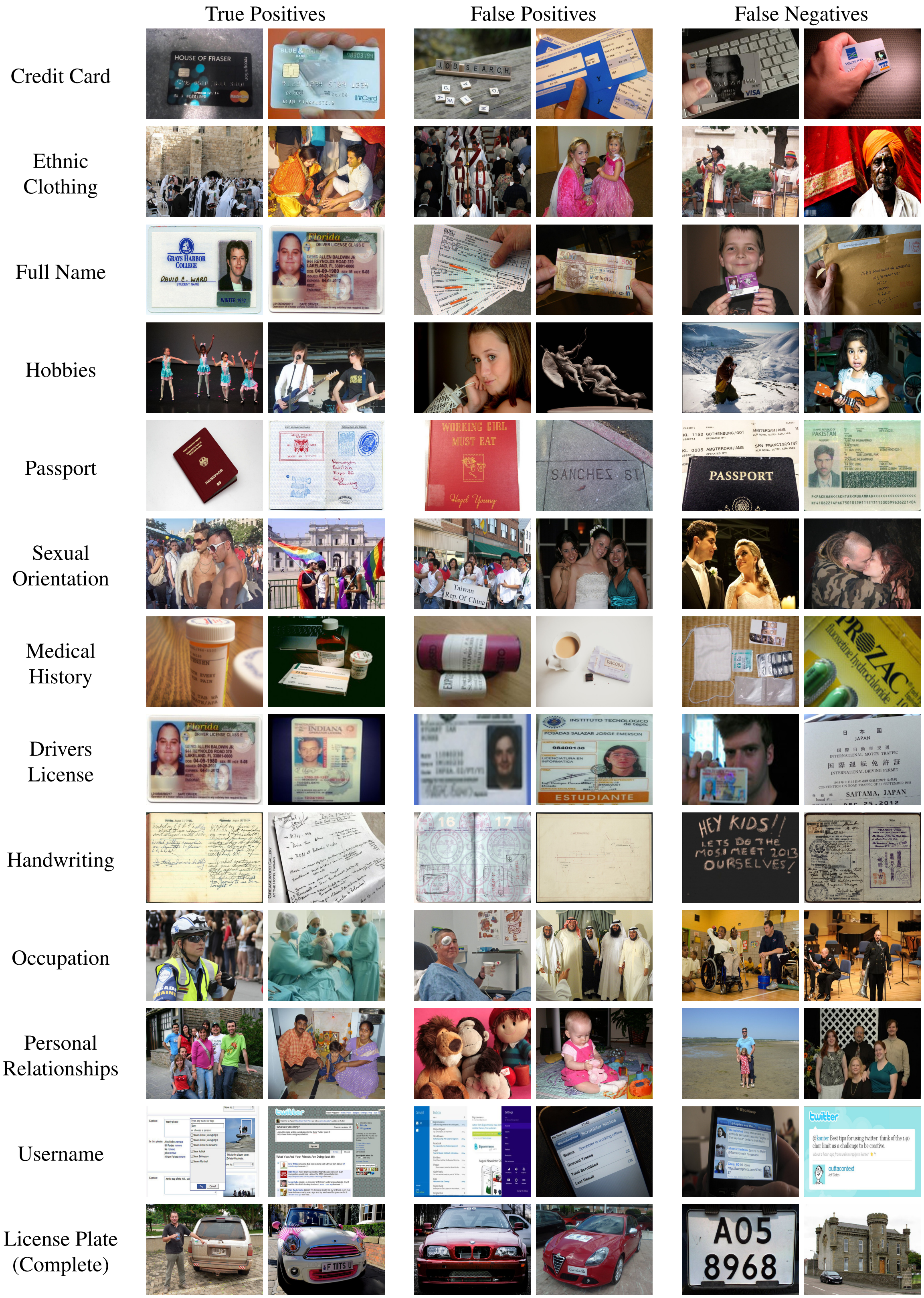


Figure 3: Additional Qualitative Results of our Privacy Attribute Prediction method

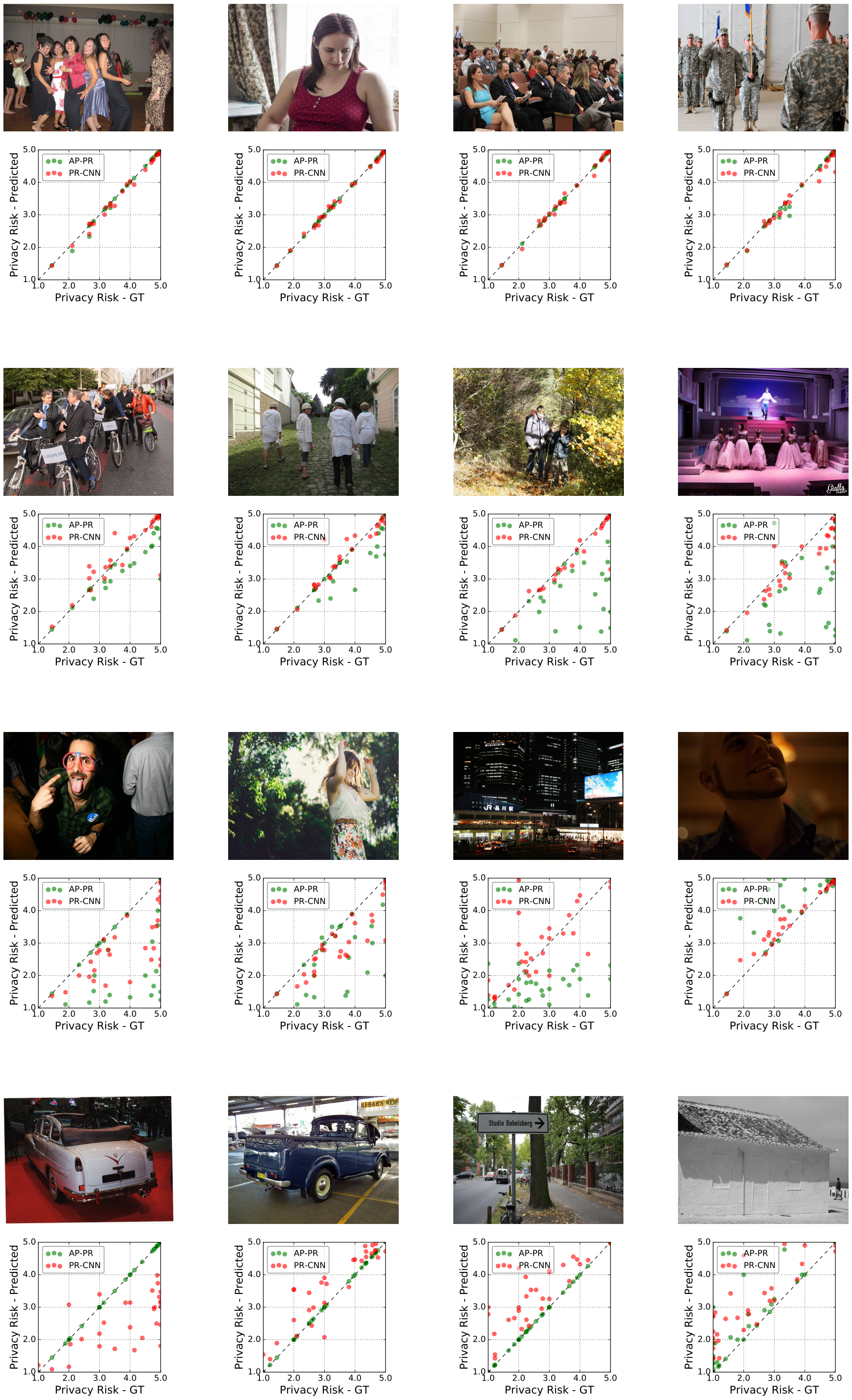


Figure 4: Qualitative results for Personalized Privacy Risk Prediction

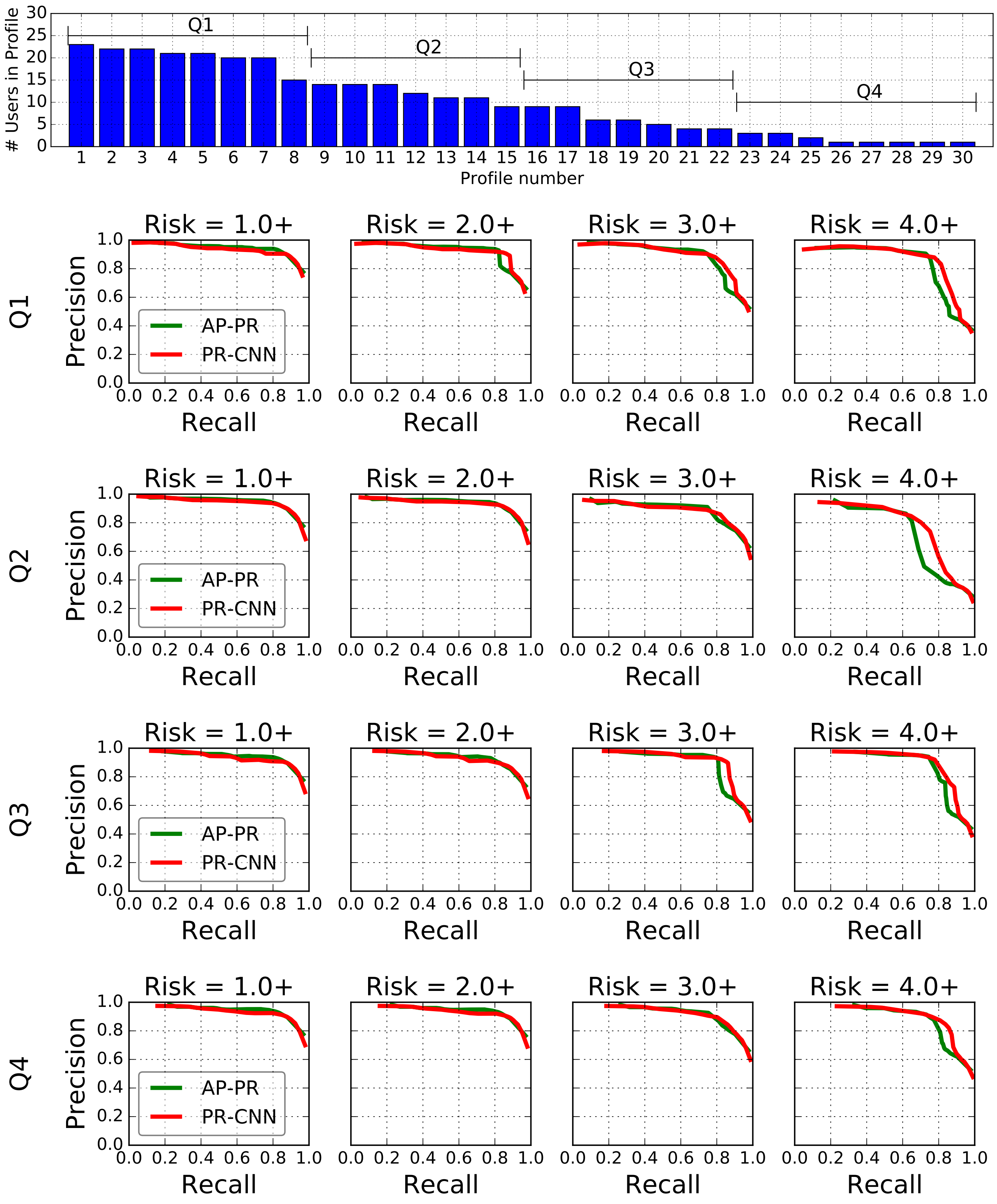


Figure 5: Precision-Recall curves when visualized over groups of user profiles

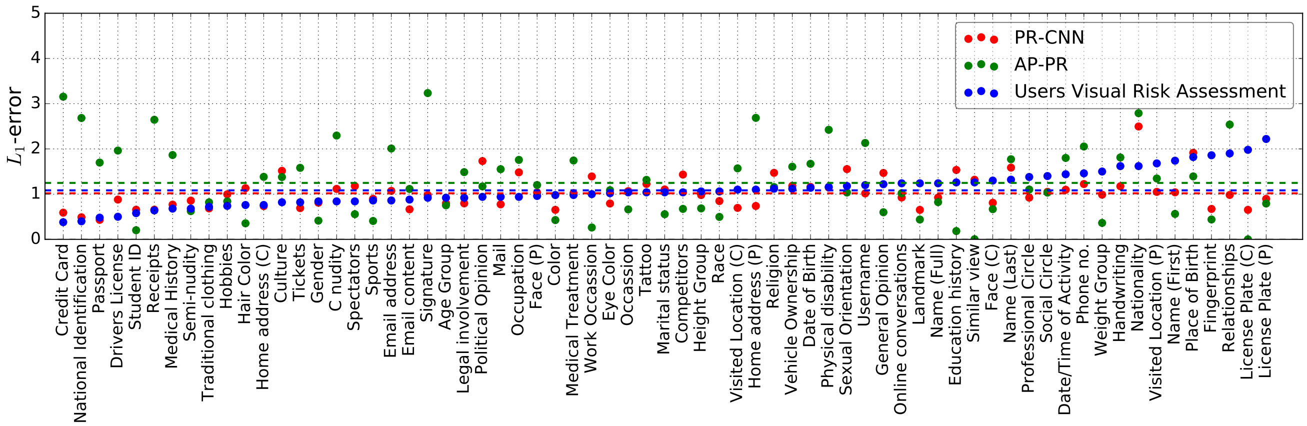


Figure 6: L_1 errors over attributes

visible license plates or family photographs depicting relationships. In these cases, PR-CNN is better at evaluating risks. (iii) Considering the attributes in which AP-PR incurs high errors (*e.g*. relationships, addresses, username, signature, credit card), we see that PR-CNN outperforms in all these cases bypassing incorrect attribute predictions.