Automatic Privacy Classification of Personal Photos

Daniel Buschek[1(✉)], Moritz Bader[1], Emanuel von Zezschwitz[1], and Alexander De Luca[1,2]

[1] Media Informatics Group, University of Munich (LMU), Munich, Germany
{daniel.buschek,emanuel.von.zezschwitz, alexander.de.luca}@ifi.lmu.de, moritz.bader@googlemail.com
[2] DFKI GmbH, Saarbrücken, Germany

Abstract. Tagging photos with privacy-related labels, such as "myself", "friends" or "public", allows users to selectively display pictures appropriate in the current situation (e.g. on the bus) or for specific groups (e.g. in a social network). However, manual labelling is time-consuming or not feasible for large collections. Therefore, we present an approach to automatically assign photos to privacy classes. We further demonstrate a study method to gather relevant image data without violating participants' privacy. In a field study with 16 participants, each user assigned 150 personal photos to self-defined privacy classes. Based on this data, we show that a machine learning approach extracting easily available metadata and visual features can assign photos to user-defined privacy classes with a mean accuracy of 79.38 %.

Keywords: Photos · Privacy · Classification · Images · Metadata

1 Introduction

Browsing a personal photo gallery in the presence of others can unintentionally reveal private content and therefore violate the user's privacy. Moreover, users may want to share their photos online on social platforms like Flickr or Facebook. To support users in revealing their photos only to intended audiences, applications need to become more privacy-aware, as recommended in related research [1, 10, 11].

To achieve this, applications should be informed about the user's intended privacy setting, ideally for each photo individually [1]. Unfortunately, manually sorting photos with respect to privacy concerns and different audiences is time-consuming or not feasible at all for large collections.

We present an approach to automatically assign personal photos to user-defined privacy classes. For example, users might define classes related to specific places, activities, events or audiences. We employ a machine learning approach to infer each photo's intended privacy class from metadata (e.g. timestamp, GPS, ISO-speed) and from visual features (e.g. based on colours, edges, occurrences of faces). Our insights enable more privacy-aware photo applications and thus support users in sharing their photos only with their intended audiences. Our contribution is twofold:

J. Abascal et al. (Eds.): INTERACT 2015, Part II, LNCS 9297, pp. 428–435, 2015.
DOI: 10.1007/978-3-319-22668-2_33

- *We describe and evaluate metadata and visual features with respect to users' own privacy classification.* In a user study ($N = 16$), participants sorted personal photos into three self-defined privacy categories. We show that a Random Forest classifier matches these user-defined categories with 79.38 % accuracy.
- *We present a privacy-respecting study method to gather relevant image data from personal photos.* We implemented a tool for users to extract image data themselves at home. Hence, they never had to reveal their photos.

2 Related Work

Ahern et al. [1] investigated privacy decisions and considerations in mobile, online photo sharing. They examined photo tags and privacy settings with a Flickr app, showing that tags related to persons and locations had the highest ratio of private photos. They derived a taxonomy of privacy considerations from their data analyses and interviews. This revealed that users' main concerns were related to influencing their own online identity and those of others. The authors concluded that applications should help users to prevent privacy-related mistakes when sharing photos online. This supports the motivation for our work. Furthermore, the results from our feature analysis match their findings regarding the influence of persons and locations on users' privacy settings.

Zerr et al. [11] targeted privacy-aware image search with a Support Vector Machine (SVM) classifier trained on publicly available photos from Flickr. They used five visual features: faces, hue histogram, edge-direction coherence vector [7], SIFT features [5], and average brightness and sharpness. Their search method achieved 0.74 F_1-Score in retrieving private versus public photos, and 0.80 when combined with title and tags as textual features. They did not report classification accuracy.

They also proposed an alert system to warn users when uploading potentially sensitive photos [10]. However, in both projects [10, 11], the "private" photos had been published on the web by Flickr users. Hence, these pictures were not considered private by their owners. They were later tagged as private by others via a community game. In contrast, we classify personal photos not shared on the web. We employ similar visual features, but do not use textual annotations. We further include metadata features, such as timestamps and GPS locations.

Klemperer et al. [3] repurposed existing organisational image tags for access control in photo sharing. They concluded that: "It may be possible to additionally aid users with […] automated tag generation". This motivates our idea of adding new privacy tags with automatic privacy classification into user-defined classes. Klemperer et al. employed decision trees to generate access rules from tags, while we employ them to add tags from metadata and visual features. Their participants tagged photos in the lab, whereas ours sorted their photos at home. Since we aim to tag photos automatically, we are not interested in observing users, but rather in including photos, which users might not like to bring to a lab study. We present an evaluation concept to include such photos while respecting participants' privacy.

3 Approach

We first describe a threat model and explain how a system using our photo classification method protects the user's privacy in related scenarios. Thereafter, we describe our photo classification system in more detail.

3.1 Threat Model

When users are browsing through a photo gallery on their mobile device in the presence of others, such as friends, family members, or unknown passengers on a bus, these bystanders could (un)intentionally catch a look at the pictures on the screen, thus possibly violating the users' privacy. The user might also *want* to present some pictures to others, but the gallery could then also reveal private ones while browsing. In another scenario, the user uploads pictures to a social network, but only wants to share certain pictures with certain groups of people.

To avoid revealing private pictures to unwanted eyes, we imagine applications to allow users to create privacy classes and to switch between them, for example "myself'", "colleagues", "friends". Only pictures assigned to the current setting are then displayed. However, this requires the user to assign a privacy class to each picture - a potentially very tedious and time consuming task.

The system proposed in this paper addresses this issue by automatically assigning one of three user-defined privacy classes to each new picture.

3.2 Photo Classification System

We employ a machine learning approach to assign photos to privacy classes. Our method comprises three steps: First, we define and extract relevant features from metadata and the visual pixel data itself. Second, we train a classifier on these features, extracted from a set of training images. Finally, the trained classifier can assign new photos to privacy classes.

Features. We extracted two types of features: metadata (location, time, shot details) and visual features (faces, colours, edges). All features are described in Table 1, which also provides references to related work for in-depth descriptions.

The choice of examined features was heuristically guided by expectations regarding possible indicators for privacy. For example, certain locations and timeframes could be related to certain activities in the user's life, such as a holiday, sports training, nightlife, and so on. Moreover, faces reveal the presence of people in a photo; a single face might indicate a more private setting than a group of many faces. Additionally, long straight edges indicate man-made structures (e.g. indoors, in a city), while scenes in nature feature many short incoherent edges [2, 7]. Our best feature set, derived from analysis of our study data, indeed contains features related to location, time, edges and the number of faces.

Classifiers. We evaluated three common classification approaches to show that the described features can be suitably used by different methods. In particular, we evaluated: Random Forest (RF), Support Vector Machine (SVM), and Nearest Neighbour (NN). Random Forest performed best.

4 Field User Data Collection

We conducted a field study to collect user data and evaluate our approach: Users defined three privacy classes and manually assigned 50 personal photos to each class. They extracted features with a given application and sent us the resulting feature-file. Hence, we never saw the users' actual photos. Photos cannot be reconstructed from the described features.

Participants: We recruited 17 participants with an average age of 26 years (SD 9). 10 were female, 7 male. One participant was later excluded from analysis, since this user had taken all photos exclusively for this study, which renders the data artificial. Participants were compensated with a 15€ gift card for an online shop.

Apparatus: Users were given a simple application without a graphical interface. When executed, it extracted the relevant metadata and visual features from all photos within a provided folder and wrote them to a feature-file.

Procedure: Users participated remotely. We sent them the feature extraction application and study instructions: First, they defined their own privacy classes by creating a folder for each and naming it; they also added a privacy ranking (public: 1 to private: 5), for example "5_myself". Second, users browsed their photo collections for 50 pictures per class and copied them into the corresponding folders. Third, they ran our feature extraction application and uploaded the resulting file to our server. Users also filled in a short questionnaire. Finally, we asked them to assign 5 new photos to each of their privacy classes after a week.

5 Results

We used the scikit-learn library [6] for Python to implement and evaluate the proposed system. We report classification accuracy; the ratio of photos for which the automatic assignment matches the user's manual assignment. Accuracies were computed with 10-fold stratified cross-validation.

5.1 Feature Selection

We first evaluated classification accuracy when using each feature on its own. The classifiers' hyperparameters were optimised per feature for each user. Table 1 shows the results: Overall, time and location features performed best.

We then applied a wrapper feature subset selection approach [4] to find the best combination of features. To reduce the search space for the wrapper, we removed the least promising features - those which never appeared among the top half in at least one of three tests: single feature evaluation, ANOVA F-value-score, and feature importances with RF. We refer to the library documentation for further details and related reading [6]. Our wrapper method greedily tests feature sets with a given classifier (here we used RF, since it performed best for single features) and removes the feature for which the remaining set leads to the best classification accuracy.

Table 1. Single feature evaluation. For each feature, the table shows mean classification accuracy and standard deviation achieved with the three tested classifiers, when considering only this feature. Features are ranked by resulting maximum accuracy with any of the three tested classifiers. The last column shows for how many users this feature was extracted from the collected data (feature present in > 50 % of a user's photos). Highlighted are the features comprising the best feature set, as found with a wrapper subset selection approach.

	Feature	Description	RF		SVM		NN		Users
			mean	std	mean	std	mean	std	
1	**unix mins**	minutes from the photo's unix timestamp	73.75	14.65	70.46	16.41	69.67	17.34	16
2	**latitude**	latitude from the photo's GPS data	70.42	11.2	55.21	14.85	67.37	14.53	16
3	**longitude**	longitude from the photo's GPS data	69.79	13.37	55.42	15.97	67.17	13.99	16
4	unix days	days from the photo's unix timestamp	68.54	14.96	65.71	15.15	60.88	13.71	16
5	day-time	timestamp relative to the day from the photo's unix timestamp	63.75	14.92	57.04	13.63	61.67	13.48	16
6	address	country code and postcode derived from the photo's GPS data	63.38	14.67	61.50	14.44	55.75	15.6	16
7	**elevation**	metres above sea level from the photo's GPS data	63.00	15.55	60.17	14.07	58.96	15.16	16
8	**cal-week**	calendar week of the year from the photo's unix timestamp	62.75	13.7	60.83	14.52	56.04	12.33	16
9	hist-hue	histogram (16 bins) of the photo's hue distribution over all pixels	59.92	10.72	56.50	10.48	55.67	11.74	16
10	**month-day**	day of the month from the photo's unix timestamp	59.00	14.74	54.71	15.02	54.46	13.8	16
11	edcv	histogram (36 bins) of the angle distribution of detected edges [7]	52.92	10.19	50.37	9.95	49.17	9.69	16
12	exposure	exposure from the photo's metadata	52.52	6.76	44.67	9.17	48.14	5.38	14
13	hist-bright	histogram (16 bins) of the photo's brightness distribution over all pixels	52.29	9.83	46.71	9.5	49.33	8.96	16
14	hist-sat	histogram (16 bins) of the photo's saturation distribution over all pixels	51.87	10.11	47.17	9.84	47.29	10.45	16
15	**ISO-speed**	ISO-speed value from the photo's metadata	50.96	9.72	49.71	8.56	47.17	9.56	16
16	**imp. hue**	index of the largest bin of the hue-histogram	49.42	11.83	50.33	10.91	46.83	10.06	16
17	weekday	day of the week from the photo's metadata	49.25	11.01	48.87	11.42	45.13	10.14	16
18	**acutance-local**	edge strength (Sobel operator) of detected edge pixels (Canny edge detector)	41.33	9.06	46.37	6.65	43.46	8.97	16
19	holiday	flag indicating whether the day from the photo's metadata is a public holiday	44.10	8.29	44.26	8.75	39.28	8.2	13
20	edge-ratio	ratio of detected edge pixels (Canny edge detector) to the total number of pixels	42.58	7.22	43.38	8.3	44.17	7.23	16
21	flash	flag indicating whether the flash was used, taken from the photo's metadata	43.43	8.39	43.62	7.89	42.43	8.17	14
22	**num faces**	number of faces detected in the photo [8, 9]	42.58	6.09	42.54	5.65	42.17	5.2	16
23	edcv-ratio	ratio of coherent edges to non-coherent edges [7]	42.17	10.07	42.38	8.42	41.79	10.79	16
24	focal-len.	focal length from the photo's metadata	41.96	11.4	41.21	11.23	40.50	10.05	16
25	acutance-global	edge strength (Sobel operator) of every 9th pixel	41.71	8.8	41.96	7.23	41.13	9.05	16
26	brightness	perceived brightness (0.299R + 0.587G + 0.114B) averaged over all pixels	40.54	9.48	40.12	9.67	41.17	8.61	16
27	**resolution**	number of pixels	41.17	11.69	40.96	11.74	39.17	9.58	16
28	f-number	f-number from the photo's metadata	41.00	11.27	40.14	11.2	38.95	9.98	14
29	edcv-std-c	standard deviation of the angle of coherent edges [7]	39.29	5.79	39.79	6.22	40.08	7.18	16
30	orientation	flag indicating "portrait", "landscape" or "square"	39.96	3.42	39.79	3.66	35.71	5.07	16
31	edcv-std-n-c	standard deviation of the angle of non-coherent edges [7]	38.58	7.72	38.67	5.92	37.46	6.21	16
32	model	camera model from the photo's metadata	38.33	7.26	38.37	7.2	35.75	5.02	16

The search terminates when no further improvement can be achieved with changes to the feature set.

The best found set consists of 12 features (**highlighted** in Table 1) latitude, longitude, elevation, Unix minutes, calendar week, weekday, day of the month, important hue, ISO-speed, local acutance, number of faces, and resolution.

This set includes features from all examined dimensions: location, time, shot details, face detection, colours, and edge detection. Hence, all described dimensions were found to be relevant and complement each other for privacy classification of personal photos. The best classifier (RF) achieved 79.38 % (SD 11.00 %) classification accuracy with this optimised set.

5.2 Error Case Analysis

Although users were free to define any three privacy classes, not necessarily hierarchical ones, all created a hierarchy. Therefore, given our threat model, we can distinguish two types of errors: assigning an image to a less private class, or assigning it to a class of higher privacy. We can expect the first type to be more serious in most applications, since it possibly reveals private content to the wrong audience. The second type could cause manual correction efforts, but no privacy violations (assuming unambiguously hierarchical classes, as chosen by our participants). An analysis of serious mistakes showed that their ratio was close to 50 % for all classifiers (RF: 49.49 %, SVM: 49.00 %, NN 48.15 %). Hence, our classifiers were not biased towards serious classification errors.

5.3 Variability Over Time

To test variability in user behaviour over time, we asked users to manually assign 5 new photos to each of their privacy classes after a week. For each user, we then trained our system on all 150 photos from the first session and evaluated how well it could assign the 15 new photos. We observed a mean accuracy of 55.42 % (SD 19.82 %). While still better than random guessing (33 %), this is a clear decline from the cross-validation results within one session (79.38 %).

To further investigate this, we asked users whether their 15 new photos were specifically taken for the study. Unfortunately, half of them had taken at least some new photos for the study. This explains the vast decline in accuracy, since those pictures were likely taken artificially in a short timeframe and at the same location, rendering two of our main features mostly useless. Accordingly, when only evaluating data for the other half of users, mean accuracy again increased to 67.50 % (SD 20.39 %). We explain the remaining gap to the results within sessions - 82.00 % (SD 12.83 %) for this subset of users - with changes in users' mental models regarding their self-defined privacy classes.

5.4 Training Set Sizes

Our system requires labelled photos for training. In practice, the user thus has to manually assign some photos to each privacy class. Hence, it is interesting to evaluate how well the system performs with fewer training images. Figure 1 shows that standard deviations increase with less training data, but mean accuracy stays consistent. In conclusion, these results show that classification with 10 manually assigned images per class is on average as good as with 50 training images.

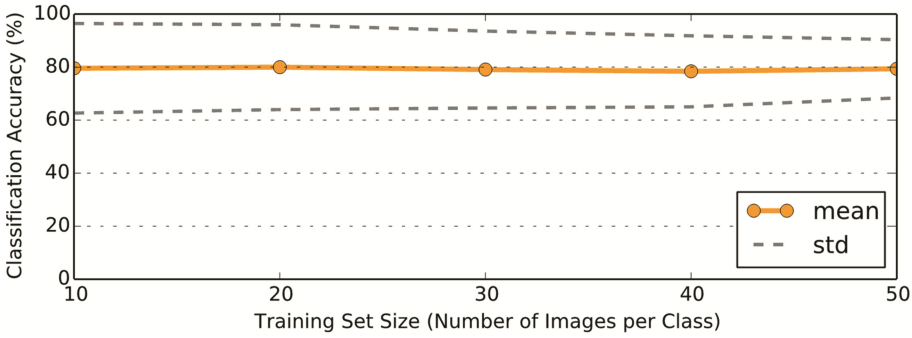

Fig. 1. Classification accuracy (RF) as a function of training set size. The plot shows that mean accuracy stays consistent when using fewer manually assigned photos for training.

5.5 Comparison to Human Reasoning

We proposed and employed a study method which respects users' privacy. Since we thus never saw their actual photos, we asked users to provide general comments on their own manual classification procedures through a questionnaire.

We analysed these comments looking for heuristics similar to our features. Users mentioned related criteria such as: presence of people (12 of 17 users); landscape and architecture (5 users); certain places or events, like "at the beach" or during a specific holiday (7 users). These considerations support the use of face detection, edge detection, time and location features, although human reasoning is of course much more complex than what can be inferred with these features. For example, regarding pictures of people, users mentioned refined criteria such as viewing angle, facial expressions, certain types of clothing, and specific individuals. Nevertheless, users' comments suggest that metadata and visual features can to some extent capture relevant aspects of human reasoning regarding privacy classes of personal photos.

6 Limitations

Extracting simple metadata and visual features, a Random Forest classifier achieved 79.38 % classification accuracy. Half of the remaining errors were assignments to less private classes. Hence, about 10 % of all automatically assigned photos could potentially violate the user's privacy, assuming hierarchical classes as created by our users. To address this issue, the automatic assignment could be presented to the user for a final check (e.g. thumbnails grouped per class).

We observed lower accuracies for assigning photos a week after training. Changing mental models of privacy classes present a challenge to automatic classification. However, classifiers could be retrained with a few new manually assigned photos to adapt to the user's mental model continuously. Additionally, long-term use of applications employing privacy classes might lead to more stable user perspectives on their self-defined classes. We leave this analysis to a future study. Nevertheless, our current system still clearly outperformed random guessing after a week.

7 Conclusion and Future Work

Privacy is an important issue when browsing personal photos on a mobile device in the presence of others, or when sharing photos online. Privacy-aware photo applications need information about the privacy class of each picture. This can lead to tedious manual assignments.

Thus, we have presented a system to automatically assign personal photos to user-defined privacy classes. In a field study with 16 participants, we showed that our machine learning approach can classify users' photos with an average accuracy of 79.38 %, based on easily available metadata and visual features. In conclusion, our approach can enhance privacy-aware applications by automatically providing privacy classes for filtering photos. Possible application scenarios include privacy-aware gallery browsing on a mobile device and selective photo sharing in a social network.

Furthermore, we have described a study method to gather relevant image data (metadata and features) without revealing the actual pictures to the researchers. We expect this approach to be useful for other studies with personal photos.

We plan to implement our system in a mobile gallery application for a deployment "in the wild". Data from long-term use can then be analysed to examine changes in users' mental models over time. Further advanced computer vision techniques could be investigated to boost accuracy, for example object recognition.

References

1. Ahern, S., Eckles, D., Good, N.: Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In: CHI 2007, pp. 357–366 (2007)
2. Kim, W., Park, J., Kim, C.: A novel method for efficient indoor-outdoor image classification. J. Sig. Process. Syst. **61**(3), 251–258 (2010)
3. Klemperer, P., Liang, Y., Mazurek, M., Sleeper, M., Ur, B., Bauer, L., Cranor, L. F., et al.: Tag, you can see it! using tags for access control in photo sharing. In: CHI 2012, pp. 377–386 (2012)
4. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1–2), 273–324 (1997)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
7. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. Pattern Recogn. **31**(12), 1921–1935 (1998)
8. van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., et al.: Scikit-image: Image Processing in Python. PeerJ2:e453 (2014)
9. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR 2001, pp. I–511–I–518, vol. 1 (2001)
10. Zerr, S., Siersdorfer, S., Hare, J.: PicAlert!: a system for privacy-aware image classification and retrieval. In: CIKM 2012, pp. 2710–2712 (2012)
11. Zerr, S., Siersdorfer, S., Hare, J., Demidova, E.: Privacy-aware image classification and search. In: SIGIR 2012, pp. 35–44 (2012)