Privacy-Aware Image Classification and Search

Sergej Zerr*, Stefan Siersdorfer*, Jonathon Hare**, Elena Demidova*  
*L3S Research Center, Hannover, Germany  
{zerr,siersdorfer,demidova}@L3S.de  
**Electronics and Computer Science, University of Southampton, Southampton, UK  
jsh2@ecs.soton.ac.uk

ABSTRACT
Modern content sharing environments such as Flickr or YouTube contain a large amount of private resources such as photos showing weddings, family holidays, and private parties. These resources can be of a highly sensitive nature, disclosing many details of the users’ private sphere. In order to support users in making privacy decisions in the context of image sharing and to provide them with a better overview on privacy related visual content available on the Web, we propose techniques to automatically detect private images, and to enable privacy-oriented image search. To this end, we learn privacy classifiers trained on a large set of manually assessed Flickr photos, combining textual metadata of images with a variety of visual features. We employ the resulting classification models for specifically searching for private photos, and for diversifying query results to provide users with a better coverage of private and public content. Large-scale classification experiments reveal insights into the predictive performance of different visual and textual features, and a user evaluation of query result rankings demonstrates the viability of our approach.

Categories and Subject Descriptors
H.3.5 [Online Information Services]: Web-based services

General Terms
Algorithms, Human Factors

Keywords
image analysis, privacy, classification, diversification

1. INTRODUCTION
With increasing availability of content sharing environments such as Flickr, and YouTube, the volume of private multimedia resources publicly available on the Web has drastically increased. In particular young users often share private images about themselves, their friends and classmates without being aware of the consequences such footage may have for their future lives [4, 26]. Photo sharing users often lack awareness of privacy issues. A recent study revealed that more than 80% of the photos publicly shared by young people are of such a private nature that they would not show these images to their parents and teachers [26]. The popular Facebook platform allows its users not only to publish photos, but also to mark the names of the depicted people. In this way, even people who did not publish any compromising information, can leave discoverable footprints on the Web.

Existing sharing platforms do not support users in making adequate privacy decisions in multimedia resource sharing. On the contrary, these platforms quite often employ rather lax default configurations, and mostly require users to manually decide on privacy settings for each single resource. Given the amount of shared information this process can be tedious and error-prone. In this paper we tackle the problem of supporting users in making privacy decisions in the context of image sharing (especially for large batch uploads). To this end, we first exploit an average community notion of privacy gathered through a large-scale social annotation game. We then employ the obtained community feedback to build classification models on selected visual features and textual metadata, and apply these models to estimate adequate privacy settings for newly uploaded images and search results. Such alert systems could be directly integrated into social photo sharing applications like Flickr or Facebook, or provided as a browser plugin.

Figure 1: Top-3 search results with original Flickr ids for the query “ronaldo” (a) in the original ranking, and (b) obtained through our privacy-oriented search method (status: Oct. 2010).
A second aspect that we study in this paper is privacy-oriented search. Existing systems do not enable users to directly search for private content in a systematic way. The motivation for enabling privacy-oriented search is two-fold: First, users should be able to retrieve resources about themselves (or about their children or other relatives) published by third parties at an early stage, so that measures such as contacting owners of servers or providers can be taken. Second, the degree of privacy of the information need behind a query can be ambiguous for a search engine. For example, a user querying “ronaldo” could be interested in photos showing either professional or private life of this football star. In this paper, we use classifier outputs to conduct privacy-oriented search, which enables users to directly discover private information about a specific topic (see Figure 1 for an example). In addition, we perform privacy-based diversification of search results (i.e. retrieving a “mixture” of private and public content) to minimize the risk of user dissatisfaction in cases where queries are ambiguous with respect to the privacy aspect of the information need. The motivation for this is analogous to topic-related diversification [11]: to cover different information needs and provide an overview over the whole search result space rather than just a list of top-ranked results.

We are aware that building alarm systems for private content and enabling privacy-oriented search can be seen as contradicting goals; privacy-oriented search is not negative per se, as it can be used for retrieving private content users are comfortable to share, and, more importantly, can help with the early discovery of privacy breaches. However, as with almost every technology, it requires sensible handling and constructive usage.

The remainder of this paper is organized as follows: In Section 2 we discuss related work in the context of security and privacy in social networks, image analysis as well as classification and diversification techniques. In Section 3 we describe our data collection and labeling method which enabled us to obtain a large dataset for the privacy-based image classification. We analyze visual and textual image features in Section 4. In Section 5 we describe the classification approach, and present ranking and diversification techniques which provide an overview of the available search results taking into account their privacy as estimated by the classifier. We evaluate our techniques for privacy-based classification, ranking and diversification in Section 6. In Section 7 we conclude and show directions for future work.

2. RELATED WORK

With emergent Web 2.0 applications, various security and privacy aspects in social networks have attracted increasing research attention. There are several works studying user privacy attitudes in social networks, privacy decisions upon sharing multimedia resources, and possible risks. In their early work [12], Gross and Aquistiu study privacy settings in a large set of Facebook users, and identify privacy implications and possible risks. Lange [18] studies user behavior with respect to revealing personal information in video sharing. All of these papers point out lack of user awareness regarding exposure of aggregated contextual information arising from users’ resource sharing habits. In this work, we provide a way to automatically identify privacy-relevant content of images to increase user awareness and support privacy decisions.

There is a plethora of work dealing with the problem of establishing suitable access policies and mechanisms in social Web environments. Caminati and Ferrari [6], for example, propose collaborative privacy policies as well as techniques for enforcing these policies using cryptographic protocols and certificates. Felt and Evans [9] suggest to limit access to parts of the social graph and to certain user attributes. Squicciarini et al. [27] introduce privacy mechanisms in social web environments where the resources might be owned by several users. In [3], the authors discuss the problem of defining fine-grained access control policies based on tags and linked data. The user can, for instance, create a policy to specify that photos annotated with specific tags like “party” can only be accessed by the friends specified in the user’s Friend of a Friend (FOAF) profile. In this paper, we do not focus on access mechanisms or policies. Instead, we concentrate on the automatic identification of private resources (more specifically, photos) and on privacy-oriented search.

Vyas et al. [32] utilize social annotations (i.e., tags) to predict privacy preferences of individual users and automatically derive personalized policies for shared content. These policies are derived based on a semantic analysis of tags, similarity of users in groups, and a manually defined privacy profile of the user. Ahern et al. [2] study the effectiveness of tags as well as location information for predicting privacy settings of photos. To this end, tags are manually classified into several categories such as Person, Location, Place, Object, Event, and Activity. In comparison, we do not require these manual steps, and, in addition, make use of the image content which allows us to utilize visual features to determine adequate privacy settings even for resources that do not have any associated tags.

Analysis of visual and textual image (meta-)data is applied to tackle a variety of problems, such as determining attractiveness [25] or quality [33] of photos, search result diversification [19], and others. Figueiredo et al. [10] analyze the quality of textual features available in Web 2.0 systems and their usefulness for classification. In comparison to these works, we study textual and visual features in the context of privacy-oriented classification and search.

Classification is a well-studied area with a variety of probabilistic and discriminative models [7]. The popular SVM-light software package [15], for instance, provides various kinds of parameterizations and variations of classification techniques (e.g., binary classification, SVM regression and ranking, transductive SVMs, etc.). In this paper we apply classification techniques in a novel context to automatically determine image privacy.

There is a large body of work on diversification of document search results and adaptation of evaluation schemes in this context - see e.g. [5, 8, 11]. Several of these papers perform diversification of search results as a post-processing or re-ranking step of document retrieval. These methods first retrieve the relevant search results, and then filter or re-order result lists to achieve diversification. In this paper, we do not deal with topic-related search result diversification but, instead, adapt and apply diversification techniques for privacy-oriented search.

In comparison to previous work, we are the first to analyze various visual features and textual annotations in the context of image privacy, and to conduct a study of their applicability in the context of privacy-oriented classification and search.
3. DATA

An individual's notion of privacy is prone to continuous change. For example, something a 16 year old youth may not consider sensitive may later lead to a rejection of his job application. Therefore, independent opinions can be very useful in order to support the user for adequate privacy judgment, and we build on an average community notion of privacy collected in this study. Our intent is to automatically suggest images with potentially sensitive content (and leave the final decision to the user), rather than images that a particular person would not share with the world.

In order to obtain an appropriate dataset with labeled private and public image examples, we performed a user study in which we asked external viewers to judge the privacy of the photos available online. To this end, we crawled 90,000 images from Flickr¹, using the “most recently uploaded” option to gather photos uploaded in the time period from January to April 2010. As we planned to investigate whether textual annotations can help us automatically selecting suitable privacy settings for an image, we only crawled images annotated with at least five English tags.

Of course, Flickr does not provide access to photos explicitly declared as private by other users; therefore, in this work we focus on private images that are publicly available on the Web. However, we think that classification and feature engineering techniques employed in this paper are general enough to be also applied to learn classification models from hidden private content.

As labeling a large-scale image dataset requires considerable manual effort, we conducted the user study as an annotation game (cf. Figure 2). Social annotation games are popular tools for solving tasks that are difficult for machines, and easy for human users, such as image labeling [31]. At each step of the game we presented five photos to a participant of the study. For each photo, the participants had to decide if, in their opinion, the photos belonged to the private sphere of the photographer. To this end, we asked the participants to imagine that images presented to them were photos they took with their own cameras, and mark these images as “private”, “public”, or “undecidable”. We provided the following guidance for selecting the label: “Private are photos which have to do with the private sphere (like self portraits, family, friends, your home) or contain objects that you would not share with the entire world (like a private email). The rest is public. In case no decision can be made, the picture should be marked as undecidable.” If the user could not decide at first glance, she could choose to display metadata of the photo or use a link to see the photo in its original context on Flickr. To obtain sufficient evidence for the privacy label assignments, each photo was shown to at least two different users. In case of a disagreement, the photo was queued to be shown to additional users. Users gathered points for each classified photo. As incentive for providing accurate answers, credit was not just given for the number of labeled images, but also for inter-user agreement. The more similar the choices of the user and the other players (their assessments, of course, not being visible to the user in advance), the higher was the score of the user. Our privacy game had an unlimited number of levels and was online over the course of two weeks. We asked our participants to complete at least one level of the game with 1000 pictures; in addition, the users could continue playing to gather more points for themselves and their teams. Over the course of the experiment, 81 users between ten and 59 years of age worked in six teams. For example, one of the teams consisted of graduate computer science students working together at a research center; other teams contained users of social platforms including facebook and some online forums². Altogether 83,820 images were annotated³. After cleaning (remove unregistered users and photos annotated by less than two different people) 37,535 images remained. This set was used for our experiments.

Our analysis revealed that in the cleaned dataset around 70% of photos were labeled as public or undecidable by all of the participating judges. This relatively high percentage is expected as we originally crawled only images publicly available on the Web. Around 13% of the images were labeled as “private” by all the judges, and 28% received “private” votes from at least one of the judges. Overall the dataset contained 4,701 images labeled as private, and 27,405 ones labeled as public by 75% of the judges, which we use as ground truth for our classification experiments in Section 6.1. Thus, depending on the voting-based threshold applied, our sample indicates that a publicly available set of images from Flickr can contain 13-28% of private images.

We computed the inter-rater agreement using the Fleiss κ-measure [13], a statistical measure of agreement between individuals for qualitative ratings. Note that according to Fleiss’ definition, κ < 0 corresponds to no agreement, κ = 0 to agreement by chance, and 0 < κ ≤ 1 to agreement beyond chance. We measured κ on a randomly selected subset of 100 images from the ground truth set labeled by 36 users, where every user rated each of those images, and obtained κ=0.6.

4. FEATURES

Image privacy is a very subjective concept. It is influenced by a number of different factors such as place, objects and persons on the photo as well as the point in time the photo was taken.

Content-based features of an image like the presence of edges, faces, or other objects may give some insights about its degree of privacy. In particular, in this work, apart from

¹All rights on pictures published in this paper are reserved to the image owners. The pictures were available on Flickr in May 2012. The pictures are supplied with their Flickr ids. These ids can be used for obtaining further legal information.

²facebook.com, forum.vingrad.ru, forum.shbnt.ru and others.

³The anonymized data as well as some of the applications based on the described classification techniques are presented at http://l3s.de/picalert

Figure 2: User interface of the privacy game. The user marks each photo as either “private”, “public” or “undecidable”. The system displays the current score of the user and the scores of the competing user teams.
textual features such as title and tags, we have selected image features which can help discriminate between natural and man-made objects/scenes (the EDCV feature) and features that can indicate the presence or absence of particular objects (SIFT). We have also selected features that measure the colors within the image (color-histograms), and a feature that measures the presence or absence of faces within the image.

Visual features and associated metadata are later used for training of models for privacy classification. Note that, similar to other machine learning scenarios, “hints” provided by individual features do not suffice for predicting the privacy character of all photos; however, in Section 6.1 we will show that the combined evidence of multiple features yields applicable results.

4.1 Visual Image Privacy Features

Digital images are internally represented as two-dimensional arrays of color pixels. This representation is difficult to use directly for classification because it is highly multidimensional and subject to noise. Instead, a process known as feature extraction is typically used to make measurements about the image content. Image features come in many forms, from the very low-level, to so-called high-level features. Low-level features are typically formed from statistical descriptions of pixels, whilst high-level features are those that have a directly attributable semantic meaning. For this study, we have selected a range of image features that could potentially be used in building a classifier that can discriminate public and private images automatically.

In the following descriptions of the visual features, we attempt to discuss theoretical motivations for the use of the particular feature, and also describe some observations from the use of the features for classification experiments within our privacy image dataset. Please note that many of the illustrations need to be viewed in color (with a quality printer or as PDF) in order to understand them fully.

4.1.1 Face Detection

We first study the hypothesis that there might be a relationship between the occurrence of faces in an image and the background of the scene with respect to privacy. Figure 3 denotes the correlation between the number of detected faces in a photo and its privacy value in our dataset.

We observe that the occurrence of faces in a picture is strongly associated with a high degree of privacy although a considerable number of faces also can be found in public images. The detection of faces within an image is a high-level feature extraction operation.

Figure 3: The distribution of human faces among private and public photos

Figure 4: Example of a public photo with a few dominant colors and a private photo.

In order to find faces in images we used an implementation of the Viola-Jones face detector [30], which uses a boosted cascade of weak classifiers to detect frontal and profile faces within the image. Two types of summary feature were generated using the face detector; the first is a single numeric feature that counts the number of faces detected within the image. The second is a variable-length vector that encodes the relative area of the bounding box around each detected face. Note that private photos do not necessarily show persons (e.g. photos of letters and private workspaces); furthermore, faces are partially not recognized from certain angles or if parts of the faces are hidden.

4.1.2 Color Histograms

Intuitively, color may be an indicator for certain types of public and private image. For example, public images of landscape scenes are very likely to have a specific color distribution.

The global color histogram of an image is one of the simplest forms of image feature. The histogram is constructed by assigning each pixel in the image to a histogram bin based on its color. The global color histogram completely discards all information about how the colors appear spatially within the image.

In this study, the histogram was calculated in the HS (Hue, Saturation) color space with each color-dimension split evenly into four segments, resulting in a histogram consisting of 16 bins.

Previous studies have shown that skin-tones across ethnicities are tightly clustered in HS space [24], so HS histograms should be particularly good at efficiently modeling skin tones and predicting the presence or absence of people, which can be observed to be a high indicator of privacy.

We also observed, as shown in Figure 4 that public photos often contain a few very strong discrete colors (resulting in a spiky histogram) whereas in private photos the colors are
distributed more uniformly. The former might correspond to artistic make up and an artificial environment, whilst the latter is typical for amateur photos. More specifically, as shown in Figure 5, which was generated from the weights of the classifiers discussed in Section 5.1, we perceive that public images tend to contain more fully saturated colors (vibrant colors), whilst the colors in private images tend to be more desaturated. We also observed that black and white photos correspond to private portraits in most cases.

### 4.1.3 Edge-Direction Coherence Vector

The edges within an image are a very powerful feature for discriminating between different types of scene, and thus might be useful for privacy classification.

In particular it has been shown previously that edge-based features can be a particularly good discriminator of indoor and outdoor images [16]. Images taken in artificial environments tend to be dominated by strong, straight, near-vertical lines, whilst those of natural environment tend to predominantly contain shorter and weaker edges in all directions [29]. The check of the top 100 positive and negative images, correctly classified using the edge feature revealed that just 20% of private photos were taken outdoor whereas 11% of public ones were taken indoor.

The edge-direction coherence vector (EDCV) [29] consists of a pair of histograms that encode the lengths and directions of edges within the image. The first histogram encodes coherent edges, whilst the second histogram encodes incoherent edges. Coherent edges are edges within the image that are relatively straight and long, whilst incoherent edges are short and/or non-straight. Each bin of the histograms represents a small range of edge directions, whilst the magnitude of the bin is proportional to the summed lengths of all the coherent (or incoherent) edges with that particular direction in the image. We implemented the algorithm described in [28], where edges and their directions are found using the Canny edge detector. Coherent edges are those that deviate in direction by less than five degrees over their length, and have a pixel length greater than 0.002% of the image area. In terms of privacy classification, the EDCV feature of public images depicting landscape scenes is likely to be dominated by incoherent edges. Private images are much more likely to contain a mixture of both coherent and incoherent edges from a mixture of the presence of people (mostly incoherent) and the background (coherent).

This is illustrated in Figure 6 where the two histograms are shown as one larger histogram with the incoherent vector on the left and coherent vector on the right. Public images of cityscapes, that contain some natural features (i.e. trees or people) would also contain a mix of both coherent and incoherent edges, however, there might be some changes in the distribution of these edges with respect to their orientation.

### 4.1.4 Quantized SIFT Features

Private and public photos typically tend to be taken in specific contexts. For example, pictures can be taken in public places like stadiums, supermarkets and airports, or in private places like home, car, or garden. Accordingly the objects contained in a photo, like sport equipment, furniture, human and animal body parts could be different and thus give us some insights about an image’s privacy. Examples of features dominating the classification decision using the machine learning model described in Section 5.1 are visualized in Figure 7. Recent advancements in low-level image description have led to approaches that are based on the detection and description of locally interesting or salient regions depicting objects and textured regions. Many region detection and description approaches have been proposed, however, the SIFT [20] descriptor remains the most popular descriptor currently. The SIFT descriptor is a 128 dimensional vector that describes local texture. Because the application of an interest point detector can lead to a variable number of SIFT features for different images, a popular technique for using SIFT features for classification is to quantize the features into a small vocabulary of “visual terms”, and to create sparse, fixed length vectors for each image that encode the number of occurrences of each type of visual term. In this study, we detected interest regions in each image by detecting peaks in a difference-of-Gaussian pyramid [20]. Approximate k-means [22] was used to learn a 12000 term codebook for vector quantization from a random sample of one million SIFT features from the dataset.

---

**Figure 6:** Example of a public photo dominated by incoherent edges and a private photo of a working place with a mix of coherent and incoherent edges.

**Figure 7:** Examples of public and private photos. The most dominating features are surrounded with boxes; the brightness of the boxes is proportional to the strength of the features.

**Figure 8:** Some discriminative public and private SIFT features along with example snippets from the original images.

**Figure 9:** Top-30 discriminative public and private SIFT features
SIFT-based features are commonly used for generic object recognition because similar combinations of SIFT features will often occur in different images of the same class of object. As public and private images often have very different types of objects depicted in them, SIFT-based features have the potential to be a powerful discriminator. Figure 8 shows some examples of the discriminative public and private visual terms within the context of an image from which the term occurs. In order to obtain the most discriminative features, we computed the Mutual Information (MI) measure [21] from information theory, which measures how much the joint distribution of features (visual terms) and categories deviates from a hypothetical distribution in which features and categories (“private” and “public” in our case) are independent of each other. Figure 9 visualizes the top-30 discriminative SIFT features for public and private images. Typically, SIFT features correlated to public images are observed to be of highly textured regions, whilst those corresponding to private images tend to depict more linear features such as edges. More specifically, as can be seen from Figure 8, SIFT features corresponding to public features tend to come from regions containing text and symbolic shapes, and also from patterns like in the case of an image containing flowers. Conversely, many SIFT features associated with private features correspond to human body parts or items like furniture, typically occurring in private environments.

4.1.5 Brightness and Sharpness

We also studied features like brightness and sharpness that have been shown to be an indicator of photo attractiveness [25]. In the same study, private amateur photos were also not found very attractive by a community of users. Our brightness feature is a single valued number that is calculated by averaging the luminance values of all the pixels in the image.

4.2 Textual Image Privacy Features

In addition to visual features, the textual annotation of images available in Web 2.0 folksonomies such as Flickr can provide additional clues on the privacy of photos. This holds partly due to correlations of topics with privacy-related image content.

Table 1 shows the top-50 stemmed tags automatically extracted from our labeled data set of private and public photos described in Section 3 using MI in a similar way as in Section 4.1.4. Obviously the majority of tags used in images falling into the private category describe personal concepts like family (babi, famili, child), emotions and sentiment (happi, love, beauti), and concepts related to the human body (hair, face, eye), whereas tags in the public category mostly refer to nature motives (snow, sunset, tree), architecture (bridg, build, architect), and inanimate objects (car, rock). This result indicates that privacy decisions can be correlated with specific terms found in tag annotations, and illustrates the potential merit of additional textual metadata for privacy prediction tasks.

5. PRIVACY EXPLORER

In this section we describe the Privacy Explorer system, which provides two types of privacy-based search mechanisms: 1) privacy-oriented search retrieving the most private search results, and 2) privacy-based search result diversification providing an overview of search results with varying degree of privacy.

5.1 Classification Model

In order to predict the privacy of photos we use a supervised learning paradigm which is based on training items (photos in our case) that need to be provided for each category. Both training and test items, which are later given to the classifier, are represented as multi dimensional feature vectors. These vectors can be constructed using tf or tf·idf weights of tags or titles, and the visual features described above. Photos labeled as “private” or “public” are used to train a classification model, using probabilistic or discriminative models (e.g., SVMs).

In our classification experiments we use linear support vector machines (SVMs) classifiers. SVMs construct a hyperplane \( \omega \cdot x + b = 0 \) that separates the set of positive training examples (photos manually tagged as “private” in our case) from a set of negative examples (photos tagged as “public”) with maximum margin. SVMs have been shown to perform very well for various classification tasks [21].

For combining multiple features, a normalization step for the input vectors is needed, since the ranges of absolute values for different features can vary and are not directly comparable. One possible normalization technique is to employ a trained sigmoid function which maps feature values from arbitrary range into the range \([0, 1]\). After this transformation the dimensions of the different input vectors are in the same range and can be combined. SVMs provide as output a value which determines the distance of a new test image from the separating hyper plane. We used Platt’s method [23] to transform this output into a probability value.
5.2 Privacy-Oriented Search

In order to enable users to get an overview of the most private images relevant to a query, privacy-oriented search retrieves a set of relevant images, and re-orders them according to descending degree of privacy (see Figure 1 in the Introduction). In order to create a list of images ranked by privacy, we can estimate the likelihood of image privacy using the output of the SVM classifier trained on a set of images labeled as “public” or “private” by the users. We described the training process of the classifier and feature selection in Section 4.

5.3 Privacy-Based Diversification

A privacy-aware image search engine should provide an overview of the available search results, taking into account their privacy. For example, for a query “wedding make-up pictures”, users might want to retrieve some professional images from wedding catalogs, examples of private wedding photos, or recent wedding photos of the user’s personal friends available on Web 2.0 platforms. In contrast to Web image search, which focuses on a few relevant results, privacy-based diversification can minimize the risk of user’s dissatisfaction in such scenarios. In what follows, we describe a framework to handle diversity in image search with respect to the privacy dimension. This framework enables us to obtain result sets of images covering a broad range of privacy degrees from publicly available images to private photos of personal friends.

5.3.1 Diversification Quality and Estimates

The relevance of a search result is typically evaluated using the standard nDCG measure [14] that accumulates the gain of each individual search result throughout the result set. \(\alpha\)-nDCG [8] is a standard evaluation measure to balance relevance of search results with their novelty. \(\alpha\)-nDCG views a search result (i.e., an image or a document) as a set of mutually independent binary information nuggets. Each nugget represents the presence of a different aspect of the user’s information need, such as a fact, or topicality of the document [8]. Compared to the standard nDCG measure, the computation of the gain \(G[k]\) of a search result at rank \(k\) in \(\alpha\)-nDCG is extended with a parameter \(\alpha \in (0, 1]\) to trade off relevance and novelty of a search result:

\[
G[k] = \sum_{n_i \in N} J(d_k, n_i) \times (1 - \alpha)\sum_{j=1}^{k-1} J(d_j, n_i),
\]

where \(N = \{n_1, \ldots, n_m\}\) is the space of the possible information nuggets, \(J(d_k, n_i)\) is a binary user judgment of containment of nugget \(n_i\) in the result \(d_k\), and \(\sum_{j=1}^{k-1} J(d_j, n_i)\) is the number of higher ranked results judged to contain nugget \(n_i\). The resulting gain \(G[k]\) is discounted by its position \(k\), and accumulated over the positions \([1, k]\) to obtain the discounted cumulative gain DCG:\n
\[
DCG[k] = \sum_{j=1}^{k} \frac{G[j]}{\log_2(1 + j)}.
\]

Finally, the DCG\([k]\) is normalized by the ideal discounted cumulative gain vector DCG\(\tilde{L}\): \(nDCG[k] = \frac{DCG[k]}{DCG(\tilde{L})}\) and used as a measure of gain.

We now describe how \(\alpha\)-nDCG, originally defined in the context of topic-based diversification, can be adapted to the privacy-based diversification. In topic-based diversification, information nuggets represent some mutually independent binary properties of documents [8]. In contrast, probability of image privacy has a continuous range of values. In our context, we interpret the notion of information nugget \(n_i \in [0, 1]\) as a numeric property of the image with \(n_i = 1\) being the most private. In our user study the values of the privacy nuggets are directly judged by the users. For the automatic diversification we used the classifier output described in Section 5.1 as estimate for the privacy nuggets values.

In this scenario we cannot assume privacy-related information nuggets in different search results to be mutually independent. Therefore, we propose a more general \(\alpha\)-nDCG-G measure that takes into account similarity of the information nuggets contained in the \(k^{th}\) result to the information nuggets contained in the higher ranked search results:

\[
G[k] = \sum_{n_i \in N} J(d_k, n_i) \times (1 - \alpha)\sum_{j=1}^{k-1} \sum_{n_y \in N} J(d_j, n_y) \times \text{sim}(n_i, n_y),
\]

where in the equation \(G[k]\) is the \(k^{th}\) element of the gain vector, \(N = \{n_1, \ldots, n_m\}\) is the space of the possible nuggets, \(J(d_k, n_i)\) is a binary judgment of containment of nugget \(n_i\) in search result \(d_k\), \(\alpha \in (0, 1]\) is a factor to trade off relevance and novelty of a search result, and \(\text{sim}(n_i, n_y)\) is the similarity between the nuggets \(n_i\) and \(n_y\).

In general, Equation 3 relaxes the nugget independence assumption of the standard \(\alpha\)-nDCG. This equation can be applied to estimate quality of a search result in presence of information nuggets with continuous range of values. As in this paper we perform diversification only with respect to the privacy dimension and image \(d_k\) only contains one particular privacy nugget \(n_{k_i}\), we can simplify Equation 3 for this particular scenario as follows:

\[
G[k] = (1 - \alpha)\sum_{j=1}^{k-1} \text{sim}(n_{k_i}, n_j),
\]

where \(n_{k_i}\) and \(n_j\) are the privacy nuggets in images \(d_k\) and \(d_j\) respectively. A possible estimate for similarity of the privacy nuggets \(n_{k_i}\) and \(n_j\) is the difference in their privacy values: \(\text{sim}(n_{k_i}, n_j) = 1 - |n_{k_i} - n_j|\).

5.3.2 Diversification Algorithm

For automatic diversification, nugget privacy values \(n_{k_i}\) and \(n_j\) required for the gain computation in Equation 4 are computed with probability estimates obtained through the classification methods described in Section 5.1.

\begin{algorithm}
\caption{Proc Select Diverse Images}
\begin{algorithmic}
\Require list \(L\) of top-k images ranked by relevance, number \(r\) of diverse images to be selected.
\Ensure list \(R\) of the relevant and diverse images.
\State \(R \leftarrow \emptyset\)
\While{\(|R| < r\)}
\Do{\{less than \(r\) elements selected\}}
\State \(i^* \leftarrow \text{argmax}_{i \in L \setminus R} G[i]\)
\State \(L \leftarrow L \setminus \{i^*\}\)
\State \(R \leftarrow R \cup \{i^*\}\)
\EndWhile
\end{algorithmic}
\end{algorithm}

Let \(L\) be the list of top-k images sorted by the probability of their relevance to the user’s information need. Analogous to [1, 11] Algorithm 1 starts with the most relevant image at
the top of $L$, and selects subsequent candidates by greedily optimizing the estimated gain in Equation 4. The worst case complexity of Algorithm 1 is $O(l \times r)$, where $l$ is the number of images in $L$ and $r$ is the number of images to be selected for the result list.

6. EVALUATION

We now elaborate on the quantitative results of different experiments for classification, privacy-oriented search, and diversification using the visual and textual features motivated in the previous sections.

6.1 Classification

For our classification experiments, we used the SVM-light [15] implementation of linear support vector machines (SVMs) with default parameterization.

Our dataset described in Section 3 contained 4,701 images labeled as private, and 27,405 ones labeled as public by the top of the judges. In order to obtain a balanced set, we randomly restricted the set of public photos to a subset of 4,701 images, so that the number of private images was equal to the number of public ones. From this set, we randomly selected 60% as training data for building our classifiers, and 40% as test data, with each data set containing an equal proportion of public and private instances.

We conducted our classification experiments on balanced data sets, which is commonly done the literature (see e.g. [33, 25]) in order to capture general classifier properties independent of the a priori class probabilities on a particular dataset. Our quality measures are the precision-recall curves as well as the precision-recall break-even points for these curves. The break-even point (BEP) is the precision/recall value at the point where precision equals recall, which is equal to the $F_1$ measure, the harmonic mean of precision and recall, in that case. The results of the classification experiments for selected visual features described in Section 4 and the combination of visual and textual features are shown in Figure 10. The main observations are:

- Visual features like sharpness, brightness and color histograms that have been shown to be relevant in the context of photo attractiveness classification in previous work [25] do not achieve significant discriminative performance in our scenario.
- Experiments with object related visual features, however, presented in the second row, show a substantial improvement. The occurrence of faces in photos is an intuitive indicator for privacy, reflected by a BEP of 0.63 for the face feature. The edge-direction coherence feature performs slightly better, and achieves a BEP of 0.65. In our experiments, SIFT features outperform all of the other visual features (BEP = 0.70).
The combination of all available visual features further increases the BEP to 0.74.

The pictures we used for classification experiments, obtained good quality textual metadata (e.g., titles and at least three English tags). Thus the text features could provide a short but concise summary of the image content and result in a BEP of 0.78.

Finally, the combination of the visual and textual features leads to an additional performance boost with a BEP of 0.80, showing that textual and visual features can complement each other in the privacy classification task. However, classification with only visual features alone also produces promising results, and can be useful if no or insufficient textual annotations are available as is the case for many photos on the web.

Note, that it is possible to trade recall against precision for sensitive applications. For instance, we obtain prec=0.88 for recall=0.6, and prec=0.93 for recall=0.4 for a combination of textual and visual features; even when restricting ourselves to only visual features we still obtain practically applicable results (prec=0.82 for recall=0.6, and prec=0.89 for recall=0.4). This is useful for finding specifically strong candidates of private photos in large photo sets. In the context of photo-sharing applications this might allow for tuning the sensitivity of personal “privacy warning” systems according to individual preferences.

Although privacy is a subjective concept, our data annotation captures an aggregated community perception of privacy, which turns out to be correlated to textual and visual features in a plausible way, and can be predicted.

6.2 Ranking and Diversification

6.2.1 Query Set

In order to obtain queries for ranking and diversification experiments, we filtered a set of image-related queries from an MSN search engine query log based on their target URLs. To this end, we considered only queries containing the term “image” or “photo” in these URLs.

We manually filtered out queries related to adult content, and randomly selected 50 queries for our experiments (examples being “picture of a power plant”, or “fall pictures”). For each of these, we retrieved photos and corresponding metadata from the top-1000 Flickr search results.

To assess the quality of privacy-based ranking and diversification in a real-world scenario, we performed a user study with 30 participants. We presented the keyword queries and their top-10 results retrieved by different ranking methods to the users in random order. Users were asked to judge the privacy level for each image on a 4-point Likert scale, with 1 corresponding to “clearly public” and 4 to “clearly private”.

Apart from the modified Likert scale the user assessment was a replication of the game described in Section 3.

6.2.2 Ranking by Privacy

For each query, we computed privacy-oriented rankings as described in Section 5.2. The list of test photos in descending order of their user-assigned privacy value was considered as ground truth for our experiments. We compared the order of the automatically generated rankings using Kendall’s Tau-b [17]. We chose the Tau-b version in order to avoid a systematic advantage of our methods due to many ties produced by the high number of photos with equal user ratings.

The main observations are:

- The original Flickr ranking does not consider the privacy of the images in the search results. This was reflected by a small \( \tau_b \) value of -0.04. In contrast, our methods show a clear correlation with the user-based privacy ranking.
- Since not all of the photos in the set contained usable metadata, ranking based on text with \( \tau_b = 0.26 \) is outperformed by ranking using visual features \( \tau_b = 0.31 \).
- Finally the combination of textual and visual features provides the best ranking performance \( \tau_b = 0.33 \). However, ranking with only visual features can still be useful for cases and applications where no or insufficient textual photo annotation is available.

The results of the pairwise t-test between the different ranking types confirm their statistical significance for a confidence level of 95%. We obtained good classification results with default parameters for SVM. Although the tuning options such as soft-margin parameter \( C \) or alternative kernels might lead to further improvements, this is outside of the focus of this work.

6.2.3 Diversification

We used alpha-nDCG-G in order to measure the quality of privacy-based image diversification described in Section 5.3. In our experiments we used an \( \alpha \) value of 0.5 to balance the amount of re-ranking of the original search result with respect to the privacy dimension. The (hypothetical) optimal ranking for normalizing DCG was obtained through diversifying images directly by their user-defined privacy scores.

The main result observations shown in Figure 11 are:

- Our diversification approaches clearly help to diversify Flickr query results according to the privacy dimension, and, compared to the original Flickr rank, achieve an improvement of 8.7% using textual and 9.8% using visual features.
- The combination of visual and textual features leads to a further average improvement (10.5%).

For each ranking position \( k \geq 2 \) we conducted pairwise t-tests between the different ranking methods for a confidence level of 95%. The tests between the original rank and other methods confirm the statistical result significance. The difference between the textual features and the combination of the textual and visual features is significant for \( k \geq 5 \).

\[ \text{http://research.microsoft.com/en-us/um/people/nickcr/vscd09/} \]
7. CONCLUSIONS AND FUTURE WORK

In this paper we applied classification using various visual and textual features to estimate the degree of privacy of images. Classification models were trained on a large-scale dataset with privacy assignments obtained through a social annotation game. We made use of classifier outputs to compute ranked lists of private images as well as search results diversified according to the privacy dimension. Our classification and ranking experiments show the best performance for a hybrid combination of textual and visual information. However, the approach of using only visual features shows applicable results as well and can be applied in scenarios where no textual annotation is available (e.g. personal photo collections or mobile phone pictures).

Regarding future work, we aim to study additional features in the privacy context such as color-SIFT, or applying and testing various alternative machine learning approaches for classification and ranking. For gathering additional training data, techniques like active learning algorithms, which take intermediate user feedback into account, might help to further optimize the choice of sample images presented to human assessors. As the perception of “privacy” is highly subjective and user-dependent, we would like to study recommender mechanisms, such as collaborative filtering, to account for individual user preferences, and to provide personalized results.

8. ACKNOWLEDGMENTS

This work is partly funded by the European Commission under the grant agreement 270239 (ARCOMEM), 248984 (GLOCAL), and 231126 (LivingKnowledge) as well as by the NTH School for IT Ecosystems.

9. REFERENCES