Analyzing images' privacy for the modern Web

Anna Cinzia Squicciarini
College of Information Sciences and Technology
Pennsylvania State University
asquicciarini@ist.psu.edu
University Park, PA
United States

Cornelia Caragea
Computer Science and Engineering
University of North Texas
ccaragea@unt.edu
Denton, Tx
United States

Rahul Balakavi
Computer Science and Engineering
Pennsylvania State University
rmb347@psu.edu
University Park, PA
United States

ABSTRACT

Images are now one of the most common forms of content shared in online user-contributed sites and social Web 2.0 applications. In this paper, we present an extensive study exploring privacy and sharing needs of users' uploaded images. We develop learning models to estimate adequate privacy settings for newly uploaded images, based on carefully selected image-specific features. We focus on a set of visual-content features and on tags. We identify the smallest set of features, that by themselves or combined together with others, can perform well in properly predicting the degree of sensitivity of users' images. We consider both the case of binary privacy settings (i.e. public, private), as well as the case of more complex privacy options, characterized by multiple sharing options. Our results show that with few carefully selected features, one may achieve extremely high accuracy, especially when high-quality tags are available.

1. INTRODUCTION

Images are now one of the most common forms of content shared in online user-contributed sites and social Web 2.0 applications. Sharing takes place both among previously established groups of known people or social circles (e.g., Google+, Flickr or Picasa), and also increasingly with people outside the user's social circles, for purposes of social discovery [6]- to help them identify new peers and learn about peers' interests and social surroundings. For example, people on Flickr or Pinterest can upload their images to find social groups that share the same interests [6, 50]. However, semantically rich images may reveal content-sensitive information [2, 38, 49]. Consider a photo of a student's 2013 New Years' public ceremony, for example. It could be shared within a Google+ circle or Flickr group, or it could be used to discover 2013 awardees. Here, the image content may not only reveal the users' location and personal habits, but may unnecessarily expose the image owner's friends and acquaintances.

Sharing images within online content sharing sites, therefore, may quickly lead to unwanted disclosure and privacy violations [10, 2, 5]. Malicious attackers can take advantage of these unnecessary leaks to launch context-aware attacks or even impersonation attacks [22], as demonstrated by a proliferating number of cases of privacy abuse and unwarranted access.

To this date, online image privacy has not been thoroughly studied, as most of recent work has been devoted to protecting generic textual users' online personal data, with no emphasis on the unique privacy challenges associated with image sharing [25, 20]. Further, work on image analysis has not considered issues of privacy, but focused on semantic meaning of images or similarity analysis for retrieval and classification (e.g. [11, 36, 29, 17, 16]). Only some recent work has started to explore simple classification models of image privacy [38, 49].

In this work, we carry out an extensive study aiming at exploring the main privacy and sharing needs of users' uploaded images. Our goal is to develop learning models to estimate adequate privacy settings for newly uploaded images, based on carefully selected image-specific features. To achieve this goal without introducing unneeded overhead and data processing, we focus on two types of image features: visual-content images and metadata. Within these feature types, we aim to identify the smallest set of features, that by themselves or combined together with others, can perform well in properly predicting the degree of sensitivity of users' images.

To achieve this goal, we develop and contrast various learning models, that combine an increasingly large number of features using both combined and ensemble classification methods. Our analysis shows some interesting performance variability among all the analyzed features, demonstrating that while models for images' privacy can be well captured using a large amount of features, only some of them have a high discriminative power.

We specifically identified SIFT (Scale-Invariant Feature Transformation) and TAGS (image metadata) as the best performing features in a variety of classification models. We achieve a prediction accuracy of 90% and a Break Even Point (BEP) of 0.89 using these features in combination.

Furthermore, we analyze privacy needs of images on a multi-level scale, consistent with current privacy options offered by most popular Web 2.0 sharing sites and applications. We adopt a five-level privacy model, where image disclosure can range from open access to disclosure to the owner only. In addition to the five-privacy levels, we also introduce new degrees of disclosure for each image, to model

the different ways an image can be made available online. These degrees of disclosure are *View, Comment, Download*. According to this multi-level, multi-class privacy framework, we build models to estimate adequate privacy settings, using the best combination of features obtained in our privacy prediction models for binary classification. In these new models, we account for the inter-relations between different privacy classes. An example for such an inter-relation is the following: an image can be downloadable only if it can be viewed. To model these inherent inter-relations, we used Chained classifier [33] models, where predicted class labels are used to predict new class labels. Our experiments confirm that these models, executed using features such as Edge, SIFT, TAGS, and their assembled combinations, consistently outperformed strong baseline models.

To the best of our knowledge, this is the first and most comprehensive study carried out to date on large-scale image privacy classification, that includes not only simple privacy classification based on binary labels, but also models for more complex, multi-facet privacy settings.

The rest of the paper is organized as follows. We discuss prior research in Section 2. In Section 3 we elaborate our problem statement, whereas in Section 4 we discuss different image-based features that we explored. In Section 5 we analyze the patterns of visual and textual features in public and private images. In Section 6, we introduce the multi-class model. We finish our analysis in Section 7, where we discuss pointers to future works and conclude the paper.

2. RELATED WORK

A number of recent studies have analyzed sharing patterns and social discovery in image sharing sites like Flickr [15, 4, 28, 50]. Among other interesting findings, scholars have determined that images are often used for self and social disclosure. In particular, tags associated with images are used to convey contextual or social information to those viewing the photo [35, 30, 14, 4], motivating our hypothesis of using metadata as one among other features for privacy extraction.

Miller and Edwards [28] further confirm that people who share their photos maintain social bonds through tagging together with online messaging, commenting, etc. They also identify two different types of users (normal and power users), indicating the importance of interpersonal differences, users may have different levels of privacy concerns depending on their individual level of privacy awareness and the image content.

Ahern et al. [2] analyzed effectiveness of tags as well as location information in predicting privacy settings of the photos. Further, they conducted an early study to establish whether content (as expressed by image descriptors) is relevant to image's privacy settings. Based on their user studies, content is one of the discriminatory factors affecting image privacy, especially for images depicting people. This supports the core idea underlying our work: that particular categories of image content are pivotal in establishing users' images sharing decisions. Jones and colleagues [23] later reinforced the role of privacy-relevant image concepts. For instance, they determined that people are more reluctant to share photos capturing social relationships than photos taken for functional purposes; certain settings such as work, bars, concerts cause users to share less. These studies also revealed significant individual differences within the same type of image, based on some explanatory variables relating to the identity of the contacts and the context of photo capture, providing insights into the need for customized, subjective models for privacy patterns. Zerr and colleagues recently developed PiCAlert [49], which carries out content analysis for image private search and detection of private images.

Along the same theme, Besner et al. [5] pointed out that users want to regain control over their shared content but meanwhile, they feel that configuring proper privacy settings for each image is a burden. Similarly, related work suggests sharing decisions may be governed by the difficulty of setting and updating policies, reinforcing the idea that users must be able to easily set up access control policies [3, 42, 13, 38, 20, 25]. Some notable recent efforts to tackle these problems have been conducted in the context of tag-based access control policies for images [46, 24, 42], showing some initial success in tying tags with access control rules. However, the scarcity of tags for many online images [40], and the workload associated with user-defined tags precludes accurate analysis of the images' sensitivity based on this dimension only. Other work [19, 13, 3, 8, 7, 20] has focused on generic users' profile elements, and typically leveraged social context, rather users' individual content-specific patterns.

A loosely related body of work is on recommendation of tags for social discovery [35, 30, 47, 14] and for image classification [34, 48, 14] in photo sharing websites like Flickr. In these works, the typical approach is for authors to firstly collect adequate images and then classify images according to visual and context-related features. After users upload their images, the server extracts features, then classifies and recommends relevant tags and groups.

Finally, there is a large body of work on image content analysis, for classification and interpretation (e.g., [11, 36, 29, 41, 51, 43, 18]), retrieval ([17, 16, 21, 12] are just some examples), and photo ranking [39, 45], also in the context of online photo sharing sites, such as Flickr [34, 48, 14, 40, 31, 32, 15]. This previous work is useful in identifying meaningful content-based features for effective image content extraction, discussed in Section 4.

3. PROBLEM STATEMENT

The objective of our work is to explore the main characteristics of users' uploaded images, and leverage them to automatically determine images adequate privacy settings. Our goal is two-fold: (1) We aim to identify a variety of features that can be informative in profiling images' privacy needs. (2) Among the identified features, wish to determine the smallest set of features, that by themselves or combined together with others, can perform well in properly defining the degree of sensitivity of users' images.

To meet these goals, we analyze images based on visual content features and their associated metadata. The intuition underlying content-based features is that, as demonstrated in recent work (e.g. [5]) although privacy is a subjective decision, certain visual content is likely personal or too sensitive to be disclosed to a public online audience. Hence, we expect that certain visual elements of an image, like the presence of edges, its color, its predominant elements, or the presence of faces, may give some insights about its degree of privacy. Metadata, typically defined in terms of keywords extracted from tags or captions, can provide insights on the image's context, i.e. where it was taken, what it represents

to the labeler (e.g. the image owner) what feelings it evokes, etc.

Additional dimensions are purposely not considered for the purpose of this study. For instance, we do not consider any of the additional social networking or personal information about the photo owners and the site where the image is originally posted, as we aim to leverage to the extent possible the content carried by the image itself. Further, information about a photo poster and his online social network activities may not be available or easily accessible.

Our learning models try to address the stated goals using a blend of visual and metadata features using two alternative privacy models. First is a binary model, and accounts for the case of an image which is either to be disclosed or not (public vs. private). The second privacy model accounts for the more complex case of an image to be placed in a social networking site, where users may have more sophisticated options to choose from (i.e. should friend view the images? should they be allowed to download it? should family members be allowed to view and or comment on the image?). In this case, privacy settings will be defined by multi-option privacy privileges and various disclosure options.

4. IMAGE RELEVANT FEATURES

In this section, we discuss the image features considered for our analysis.

4.1 Visual Features

We are interested in identifying few pivotal features that can serve useful for image classification w.r.t. privacy. We next describe our selected visual features, and provide some observations from the use of the features for privacy models.

- *SIFT* or Scale Invariant Feature Transform [27]. As images privacy level may be determined by the presence of one or more specific objects, rather than the whole visual content itself (e.g. think about an image with somebody carrying a gun and the same image with the person holding flowers instead), features able to detect interesting points of images are needed. SIFT, being one of the most widely used features for image analysis in computer vision, is such a feature. It detects stable key point locations in scale space of an image. In simple terms, the SIFT approach is to take an image and transforms it into a "large collection of local feature vectors" [26]. Each of these feature vectors is invariant to any scaling, rotation or translation of the image. We extract a image profile based on the state-of-the-art model called Bag-of-visual-words (BoW) [37, 44], which can effectively evaluate the image similarity, and is widely used in image similarity search, image retrieval and even content-based social discovery [35]. The image BoW vector is first obtained by extracting the features of preferred images, then clustering them into the visual word vocabulary $\triangle$, where each element is the distinct word occurrence. Features are extracted for each image and each element of the feature vector is mapped onto one of the bag of words and once all the elements are checked, we get a sequence of numbers whose length is equal to the length of the Bag-of-visual-words . Each number represents the number of elements in the original SIFT feature vector, which have been mapped onto the corresponding visual word. As a result, an image profile is created $S = \{s_1, \ldots, s_m\}$, where s_i reflects the strength of image's preference on word w_i and m is the size of $\triangle$.

Figure 1: Facial recognition examples

- *RGB* or Red Green Blue. Images with a given color and texture patterns can be mapped into certain classes, based on what is learnt from the training set. For example, instances with a pattern of green and blue may be mapped to public images, being indicative of nature. Accordingly, we include RGB feature to extract these potentially useful patterns. RGB is a content-based feature which detects six different components from an image. These components are Red, Green, Blue, Hue, Saturation and Brightness. Values corresponding to each of the variables are extracted from an image, and each feature is encoded into a 256 byte length array. This array is serialized in sparse format where each instance corresponds to the feature vector of an image.

- *Facial recognition* Images revealing one's identity are more likely to be considered private [2], although this is subjective to the specific event and situation wherein the image was taken. Henceforth, to discriminate to the extent possible between purely public events with people and other images involving various individuals, we detect the ratio that the area of faces take in the image, to identify whether they are or not prevalent elements in the image.

  We extract facial keypoints using the FKEFaceDetector framework. Information about presence of faces is encoded in two attributes for each image similar to [49]. One attribute represents the number of faces found in the image and the second attribute indicates the amount of area occupied by faces in the image.

  The framework detects faces which are straight-up and clearly visible. In some images, though there are faces visible, due to various factors the faces couldn't be detected. Images with faces not clearly visible, images in dark background, images which show faces from acute

Tags : ocean, boy, summer, vacation, lighthouse, beach, water, boat, vintagecolors

Figure 2: Example of a Flickr image and its tags

angles are few factors which can result in faces not being detected by the API. We show two sets of images where facial recognition is successful and where it fails in Figure 1.

- *Edge Direction Coherence.* As more and more users enjoy the pervasiveness of cameras and smart devices the number of online images that include some "artistic" content (landscapes, sceneries etc) is also increasing. Hence, we would like to include a feature that can help with capturing similarities in landscape images. One such feature, that has shown useful for models on landscape images, is Edge Direction Coherence, which uses Edge Direction Histograms to define the texture in an image. The feature stores the number of edge pixels that are coherent and non-coherent. A division of coherence and non-coherence is established by a threshold on the size of the connected component edges. This feature uses 72 bins to represent coherent pixels and one bin for non-coherent pixels. After separating out non-coherent pixels, we back track from a random coherent pixel to another and check if its within 5 degrees and then update the corresponding coherent bins [41].

4.2 MetaData

Annotation of online images is now common practice, and it is used to describe images, as well as to allow image classification and search by others. Users can tag an image by adding descriptive keywords of the images content, for purposes including organization, search, description, and communication. For instance, each image in Flickr has associated one or more user-added tags. We created a feature vector for each image accordingly. We create a dictionary of words from all images in the training set such that there are no duplicates (in the dictionary). Once we have the dictionary ready, each feature vector is represented in sparse format, where an entry of the vector corresponds to a word. Each unique word that is a tag for an image, has an entry in the vector. This sparse representation allows compact feature vector for each instance, removing unnecessary information about absence of keywords which are in the dictionary and not in the image.

Accordingly, we try to correlate images by using the feature vector to capture usage of the same tags. We observe that most of the images which show similar descriptive patterns have extensive word usage which are similar. An example of tags usage in Flickr is given in Figure 2. For this image, the associated words are beach, water and ocean, which all have a high degree of similarity. Similar findings are reported for other images, where tags appear extremely useful: As we further elaborate in Section 5.2.3, tags are a predominant feature for privacy classification purposes, although acceptable results are found even in absence of available metadata .

5. PRIVATE VERSUS PUBLIC IMAGES: EMPIRICAL ANALYSIS

Our analysis includes two key steps. First, we analyze a large labeled dataset of images posted online, by means of unsupervised methods, to identify the main distinctions between private and public images. Second, we investigate privacy images classification models, taking into account the results of our clustering analysis and the features discussed in the previous section.

We employ two datasets for our analysis. For the first dataset, we took a sample of Flickr images from the PiCalert dataset [49]. The PiCalert dataset includes randomly chosen Flickr images on various subjects and different privacy sensitivity. Each image in the dataset is labeled using private and public labels by randomly selected users. We focus on about 6000 images randomly sampled from the original dataset, to include actual online images (i.e. still on the site) with associated keywords. The dataset includes public and private images in the ratio of 2:3. The second dataset, was sampled from the Visual Sentiment Ontology repository [9]. The repository has a good blend of landscape images, animals, artwork, people etc. About 4000 Flickr URLs were randomly sampled from the dataset. The associated keywords were extracted from the Flickr site directly.

Some of the privacy labels were already part of the first dataset, whereas we use crowdsourcing methods (i.e. Amazon Mechanical Turk) for labeling the additional datasets and complement existing labels with more complex settings (see Section 6).

Of course, similar to [49], the adopted privacy labels only capture an aggregated community perception of privacy, rather a highly subjective, personal representation. We argue that the provided representation is is correlated to textual and visual features in a plausible way, and can be predicted using carefully crafted classification models.

5.1 Characteristics of private and public images

To understand what makes images private or public, we first explored some of the consistent characteristics among each of these two classes, considering both visual and metadata elements.

5.1.1 *Visual differences between public and private Images*

We first explored whether there are any consistent types of images, or image content that can help define private versus public images.

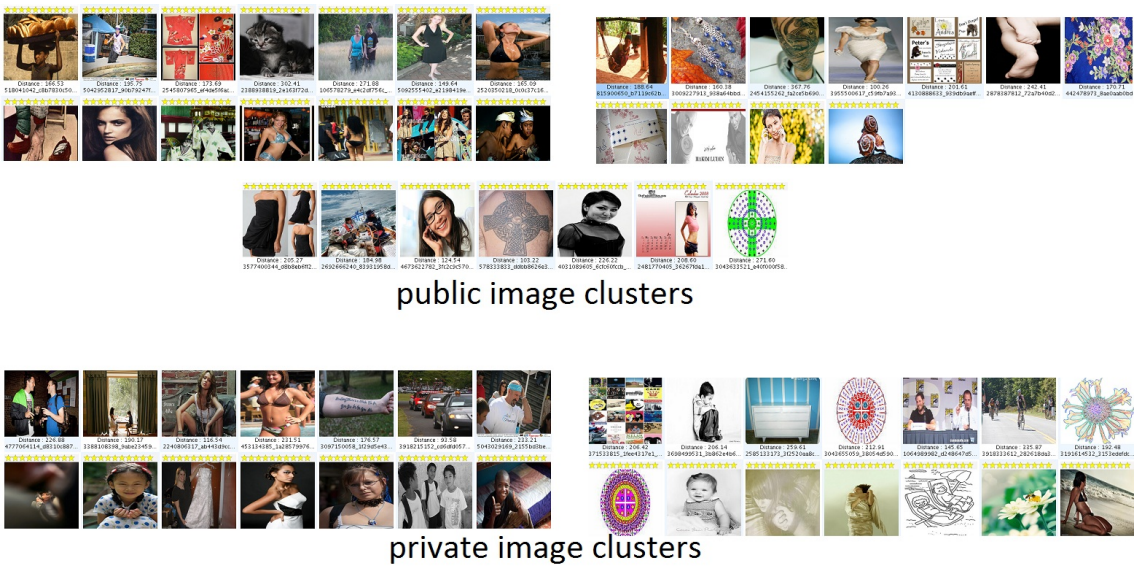

Figure 3: Public and private images

Our approach to identify these characteristics is to group images by content similarity, to explore what are the visual similarities that define the clusters. To this end, we used unsupervised learning methods. In particular, we applied Content Based Image Retrieval (CBIR) [1] method to identify clusters among public and private images, respectively. Content Based Image Retrieval uses wavelet-based color histogram analysis and enhancements provided by Haar wavelet transformation. With Color histograms, the image under consideration is divided into sub-areas and color distribution histograms are calculated for each of the divided areas. The wavelet transformation is used to capture texture patterns and local characteristics of an image. A progressive image retrieval algorithm is used for image retrieval based on similarity.

Upon running CBICR, we observed similarity patterns among images in different sets that were clustered. Figure 3 shows an excerpt of the clustered results of images. Public images are seen on the top and private images are at the bottom. Most public images belong to one of three categories. (1) Women and Children, (2) Symbols, abstract, and black and white images, (3) Artwork. Private images could be mainly grouped into (1) Women and children, (2) Sketches.

As a first observation, we note that not all images of people are private. Our clusters show that images indicative of people's life events, personal stories etc. are considered equally confidential. Second, images with children or humans in general are equally classifiable as private or public, depending on the specific visual representation in the image. These observations confirm that simply considering the presence of people as the only relevant feature may not be sufficient (we provide additional results on this aspect in the next Section), and that multiple features are needed, both visually (to describe the content in the cluster classes) and through text, to provide some contextual information.

5.1.2 *Keywords patterns in public and private images*

To further our understanding of the images and their privacy needs, we analyzed the keywords associated with each image. Specifically, we enriched the dataset by adding annotations for each image in the dataset, and extracted these annotations from the Flickr's tagging systems. Each image in Flickr has associated one or more user-added tag, which we crawled directly from Flickr as it was not part of the original dataset. To obtain keyword groups reflective of the most generic topics used to tag and index the images, we clustered images based on keyword similarity.

Precisely, we performed keyword hypernym analysis over all of the keywords associated with the images [38], using Wordnet as reference dictionary. For each keyword t_i, we created a metadata vector listing the hypernyms associated with the word. After extracting all the hypernyms of all the keywords for an image, we identified a hypernym per part of speech. We identified all the nouns, verbs and adjectives in the metadata and store them as metadata vectors $\tau_{noun} = \{t_1, t_2, \ldots, t_i\}$, $\tau_{verb} = \{t_1, t_2, \ldots, t_j\}$ and $\tau_{adj} = \{t_1, t_2, \ldots, t_k\}$, where i, j and k are the total number of nouns, verbs and adjectives respectively. This selection was done by choosing the hypernym which appeared most frequently. In case of ties, we choose the word which is closest to the root or baseline.

We repeated the same procedure over different parts of speech, i.e Noun, Verb and Adjective. For example, consider a metadata vector τ = {"cousin","first steps", "baby boy"}. We find that "cousin" and "baby boy" have the same hypernym "kid", and "first steps" has a hypernym "initiative". Correspondingly, we obtain the hypernym list η = {(kid, 2), (initiative, 1)}. In this list, we select the hypernym with the highest frequency to be the representative hypernym, e.g., "kid". In case that there are more than one hypernyms with the same frequency, we consider the hypernym closest to the most relevant baseline class to be the representative hypernym. For example, if we have a hypernym list η = {(kid,

Table 1: Prominent keywords in private images

Cluster No	Keywords
1	garment, adornment, class, pattern, appearance, representation
2	letter, passage, message, picture, scripture, wittiness, recording, signaling
3	freedom, religion, event, movement, clergyman, activity, ceremonial, gathering, spirit, group, energy
4	region, appearance, segment, ground, line, metal, passage, water, structure, material
5	body, reproduction, people, happening, soul, organism, school,class, period, respiration

Table 2: Prominent keywords in public images

Cluster No	Keywords
1	season, hour, period, decade, leisure, beginning, drib.
2	phenomenon, happening, relation, passage, electricity
3	covering, vesture, case, people,appearance, adornment, lacquerware, piece, attire, beach.
4	curio, artifact, art, crockery, lceremonial, pattern, covering

2), (cousin, 2), (initiative, 1)}, we will select "kid" to be the representative hypernym since it is closest to the baseline class "kids". Once we computed the representative hypernyms for each instance, the next step was to cluster the instances based on the hypernyms. This was achieved by calculating the edit distance of each existing cluster center with a new instance and the weighted average distance is compared to a threshold value. The new instance is added to a cluster, once the edit distance between the corresponding cluster center and the instance is below the threshold. If the distance from none of the cluster centers falls below the threshold, then the new instance is added as a part of a new cluster and the instance is made the cluster center. In addition, existing clusters keep updating their cluster centers as new instances are added. A cluster center represents a noun, verb and an adjective. These parts of speech are chosen as they are the words with highest frequency amongst the instances of the cluster.

Using the above methodology, we clustered about 6000 keywords. Keywords clustering resulted in four clusters for keywords bound with private images and five clusters for keywords related to public images. On average we observed that there were around 15 hypernyms per cluster.

Table 1 shows the prominent keywords in the clusters obtained by grouping keywords of private images. Each of the five clusters being identified projects a particular aspect or a concept (clusters are numbered for convenience only).

Cluster 1 represents adornment patterns and physical features. Cluster 2 mainly includes words about writing and communication. Cluster 3 hints at religion or a religion event. Cluster 4 indicates physical structures and perceivable entities. Keywords in the final group gravitate around children and also people at large. These results are consistent with the three image types identified by clustering images per visual content. Specifically, two of the keywords clusters (labeled for convenience as 2 and 3) are consistent with the image cluster inclusive of abstract images and images about sketches, whereas the keyword cluster with keywords surrounding children is consistent with the "women and children" cluster previously identified.

Different patterns were observed for keywords of public images. As shown in Table 2, after the clusters were formed, we observed that cluster 1 mainly grouped words related to a time scale or an event that happened in the past or that is set to happen in the future. Cluster 2 has words related to a phenomenon or something that involved movement. Cluster 3 has words which described dressing style or appearance patterns. Cluster 4 described art work or objects with patterns. These patterns shed light on the themes around private and public images. Some of these patterns (e.g. artwork) were already observed while clustering the images based on visual content using Content Based Image Representation. In addition, clustering the images based on hypernyms of the associated keywords uncovered some additional descriptive patterns, like appearance related or movement related images, that were not observed through our analysis based on content-based similarity.

5.2 Image classification models

We investigate privacy images classification models with three objectives. First, we aim to compare visual features versus metadata, to understand which class of features is more informative for privacy considerations. Second, we evaluate the performance of all the individual features used to gain an understanding of which features can be more effective in discriminating private versus public images. Finally, we aim to identify the smallest combination of features that can successfully lead to highly accurate classification.

We adopt supervised learning, based on labeled images (or items) for each category. Both training and test items are represented as multi dimensional feature vectors. Labeled images are used to train classification models. We use linear support vector machines (SVMs) classifiers. SVMs construct a hyperplane that separates the set of positive training examples (photos tagged as "private") from a set of negative examples (photos tagged as "public") with maximum margin.

5.2.1 Individual Features analysis

We begin by studying the performance of individual features. We specifically evaluated the performance of classifiers trained on 4000 labeled images and evaluated them on a test set of 500 images, using one feature per model. Results are reported in Table 3. As can be seen, TAGS is by far the best performing feature. Content-based features have lower accuracy, with the worst observed for Edge Direction feature, for which accuracy is only 54.2%.

These results provide initial evidence that users' description of images' content through tags is fairly reflective of its content. Differently, visual features by themselves appear not to be sufficient for privacy classification, most likely due

Features	Accuracy	Precision	Recall	F meas.
Tags	78.2	0.791	0.782	0.784
RGB	56.6	0.576	0.548	0.553
SIFT	59	0.613	0.59	0.594
Facial Feat.	61.2	0.584	0.612	0.514
Edge Dir. Coh.	54.2	0.588	0.542	0.542

Table 3: Individual features' classification models over 4500 images

to the heterogeneous content of the images being analyzed (e.g. facial features are inaccurate on images with landscape or food only, whereas Edge Direction does not perform well on images with faces). We also observed that most pictures that had people or faces were difficult to classify. This can be attributed to the fact that, just detecting a person on the image is not sufficient to provide an exact representation of the image. For example, a beach photo with family might be regarded as a private image. But, an image shot in a similar setting and background with a celebrity or a holiday advertisement might be regarded as public. This variation is very difficult to capture accurately without the help of user-defined keywords, which contribute to add contextual information to the image.

Hence, a more sophisticated model combining some of these features needs to be carefully designed.

5.2.2 *Models combination*

We explored various combinations of supervised learning models for image privacy, using the features listed in Table 3. We were specifically interested in understanding whether the accuracy of visual features, which was low for single-feature models, could be improved by combining them into a single classifier. The intuition is that given that these visual features seem to work best on certain types of images, we aimed to test whether when combined together, they would complement one another, reaching a higher degree of accuracy.

We tested combinations of models for a fixed set of images, increasing the size of the training data, ranging from 1000 to 3700, keeping 500 as the size of test data. To combine the models, we linearly combined the vectors of data of individual features into a single vector. Our results, reported in Figure 4 for two-features combinations, show an overall consistent increase in the accuracy of prediction across most of the combinations with the increasing size of training data. The exception to this pattern is for combinations which involved Facial features, where the peak accuracy is observed when the dataset size is in the 2500-3000 interval, followed by a decrease in the prediction accuracy after the dataset size exceeded 3000 instances landmark.

In Figure 5 and Fig. 6, we report the results for models combining three and four or more features, respectively. Some interesting observations are as follows:

- All combinations, except the FACE,RGB,SIFT combo (which is much lower) have an accuracy ranging from 65% to 74%, therefore achieving sub-optimal accuracy.
- When TAGS are in any combined classifier we obtain a better model than the same model with no TAGS as a feature, validating the role of metadata to complement visual information extracted through content-specific features. This observation is valid also for two-feature models (see Figure 4), where the best accuracy is ob-

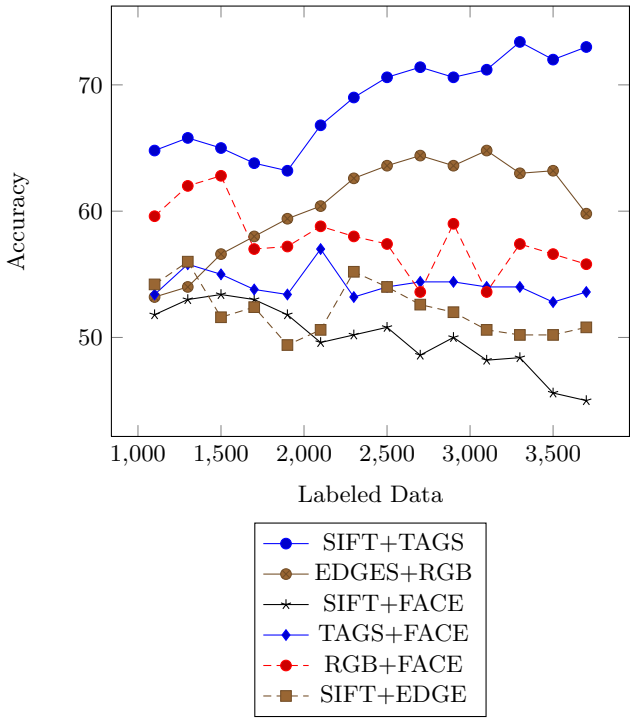

Figure 4: Accuracy analysis of combined classifiers with two features

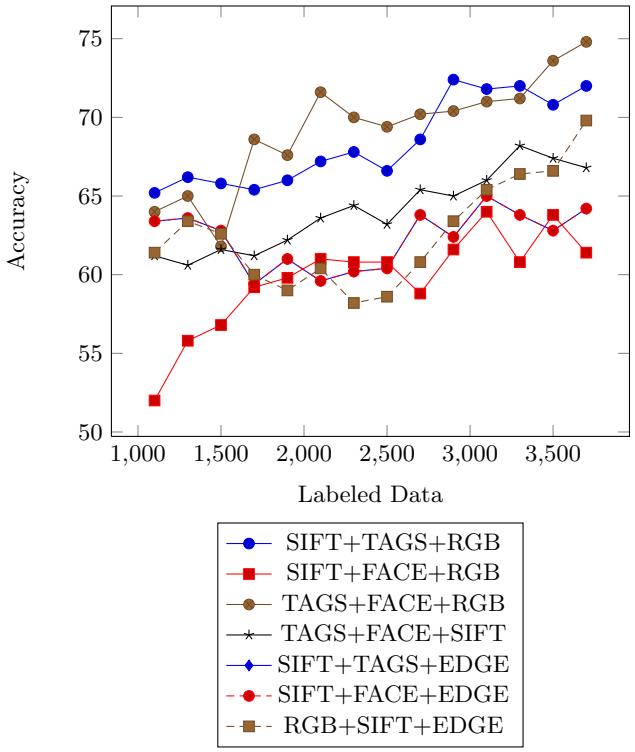

Figure 5: Accuracy analysis of combined classifiers with three features

tained with SIFT+TAGs on a labeled dataset of 3300 training instances, followed by Edge+TAGS.

- SIFT and TAGS appears to be strongest combination, for both two features and three features models. Intuitively, SIFT captured all the visual key points of the images, hence their core, discriminating visual patterns. TAGS on the other hand, gave an indication of the overall context of where the images were taken.
- When we disregard TAGS as a feature and use only visual features for prediction we reach a performance of of 70% over a dataset of size 4500 for SIFT, EDGE and RGB combined together. The combined classifier of SIFT, RGB and EDGE resulted in a BEP of 0.667.
- Finally, it appears evident that simply adding features is not always a recipe for improved accuracy: combining the visual content with metadata leads to a decreased accuracy of the metadata classifier alone (see Figure 6 for the performance of all features combination). For instance, SIFT+FACE+RGB overall perform worse than their individual features.

5.2.3 *Tags and Visual models*

In this feature study, we explore how TAGS may complement another visual feature to accurately determining adequate image privacy. We adopt a different modeling approach in an attempt to improve the prediction accuracy, and use an ensemble of classifiers, in which two classifiers are individually used to predict the outcome for a particular instance. Based on the prediction data and the confidence of prediction, the ensemble outputs a final classification result which is computed in consideration of both learning models.

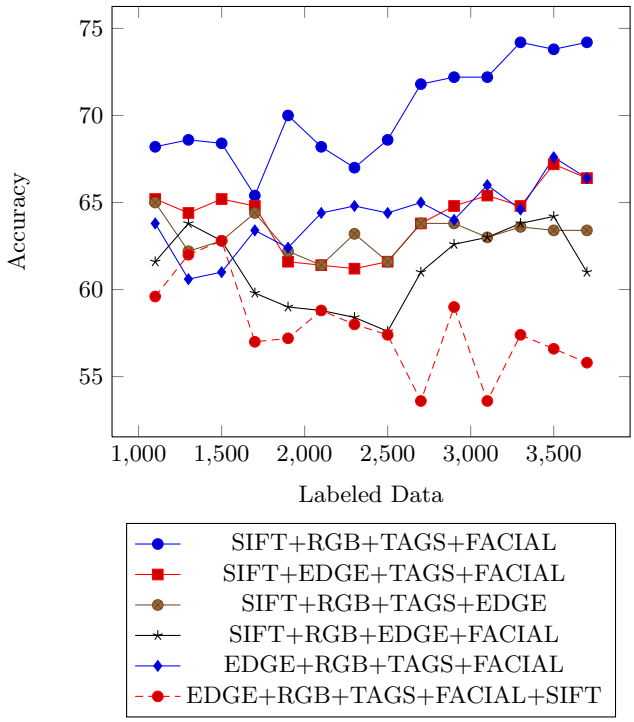

Figure 6: Accuracy analysis of combined classifiers with four or more features

In particular, in our case, an ensemble of classifiers is a collection of classifiers, each trained on a different feature type, e.g., SIFT and TAGS. The prediction of the ensemble is computed from the predictions of the individual classifiers. That is, during classification, for a new unlabeled image $\mathbf{x}_{test}$, each individual classifier returns a probability $P_j(y_i|\mathbf{x}_{test})$ that $\mathbf{x}_{test}$ belongs to a particular class $y_i \in \{private, public\}$, where $j = 1, 2$. The ensemble estimated probability, $P_{ens}(y_i|\mathbf{x}_{test})$ is obtained by:

$$P_{ens}(y_i|\mathbf{x}_{test}) = \frac{1}{2}\sum_{j=1}^{2} P_j(y_i|\mathbf{x}_{test}).$$

In experiments, we used the option `buildLogisticModel` of Weka to turn the SVM output into probabilities.

Using an ensemble of classifiers, an image that cannot be classified with good confidence by one classifier, can be helped by another classifier in the ensemble which might be more confident in classifying an image as public or private. We identified that TAGS and all the visual features can be these complementary classifiers: TAGS are collected from the keywords that a user adds to represent the image. Visual features extract the visual patterns from the image itself.

Performance metrics, obtained from a dataset of 4500 images (4000 training, 500 tests) reported in Table 4 confirm this intuition. We particularly observe a peak in the performance when SIFT and EDGE are combined together, along with ensemble of SIFT and TAGS. The latter represents the best performing combination, confirming and improving on the trend and observations made in our combination models (see Section 5.2.2). We note that (although not reported in detail) other ensemble classifiers which did not include TAGS do not reach interesting performance. We speculate that this is due to the very nature of the features, that fail to complement one another in the ensemble model.

In Figure 7 we show the precision vs recall graph for this classifier of SIFT and TAGS. Accordingly, the BEP (Break Even Point), the point where the value of precision is equal to that of recall, stands at 0.89. Compared to the Picalert framework [49] which presented a BEP of 0.80 after com-

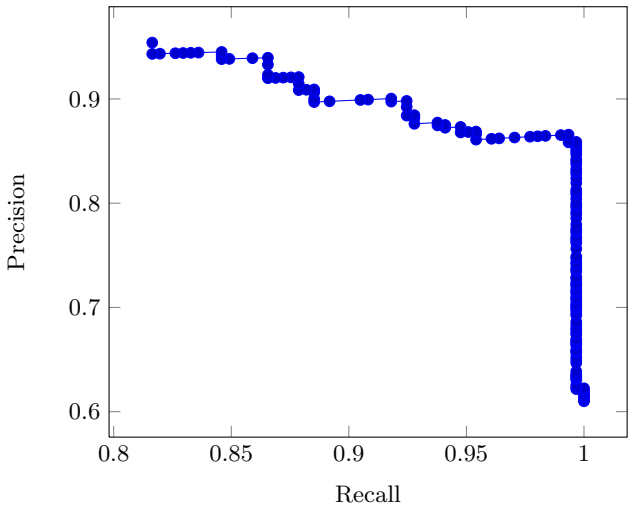

Figure 7: Precision Vs. Recall for SIFT & TAGS ensemble models

Features	Accuracy	Precision	Recall
Tags	82.4	0.859	0.824
Tags and RGB	60.4	0.605	0.604
Tags and SIFT	**90.4**	0.904	0.904
Tags and Facial	50.2	0.498	0.502
Tags and Edges	84.3	0.844	0.843

Table 4: Performance of ensemble models combining TAGS with one content feature

bining textual and a larger number of visual features, the ensembles classifier of SIFT and TAGS shows a stronger performance. In short, these results demonstrate that the ensemble of classifiers can capture different aspects of images that are important for discriminating images as private vs. public, with a small set of features. Note that these experiments also confirm the poor performance of Facial features, which achieve very low precision and recall, showing how the choice of visual features is to be carefully made.

5.2.4 *Secondary dataset experiments*

To ensure that our models were not bound to a specific dataset, we sampled a new set of images from the Visual Sentiment Ontology [9] repository. The sampling was done by randomly selecting a URL form the complete list of about 45000 images. Using the sampled set, we tested our best classifier, ensemble of SIFT and TAGS, to verify whether its performance would be similar to the performance observed on the tests carried out using the Picalert dataset. In the experiment, we varied the size of the training dataset from 1100 to 2500. Figure 8 shows the models' prediction accuracy, computed using TAGS and ensemble of TAGS and SIFT, across both (Picalert and Visual Sentiment Ontology) the datasets. We make the following observations:

- The obtained accuracy is comparable with the accuracy observed for the models against the PiCalert dataset.
- TAGS performed slightly better on the Picalert dataset, whereas the ensembles combination of SIFT and TAGS performed slightly better on the VSO dataset. This is likely due to the larger availability of high quality tags available in the first dataset. Nevertheless, we note that the final prediction accuracies across different sizes of dataset are within a range of 2-3% when we compare the results from the two datasets.
- We also observed that the pattern of increase and decrease in predication accuracy with change in size of the training set was consistent across both the datasets.

6. MULTI-OPTION PRIVACY SETTINGS

In many online sharing sites (Facebook, Flickr etc), users create a web space wherein they can control the audience visiting it, distinguishing among friends, family members or other custom-social groups, and within these groups, distinguishing the possible access privileges. Accordingly, upon analyzing image privacy using a binary classification model for "public" or "private", we investigate complex multi-label, multi-class classification models, specifically after the options offered to Flickr users, who may distinguish among multiple classes of users and sharing options (view, comment, download).

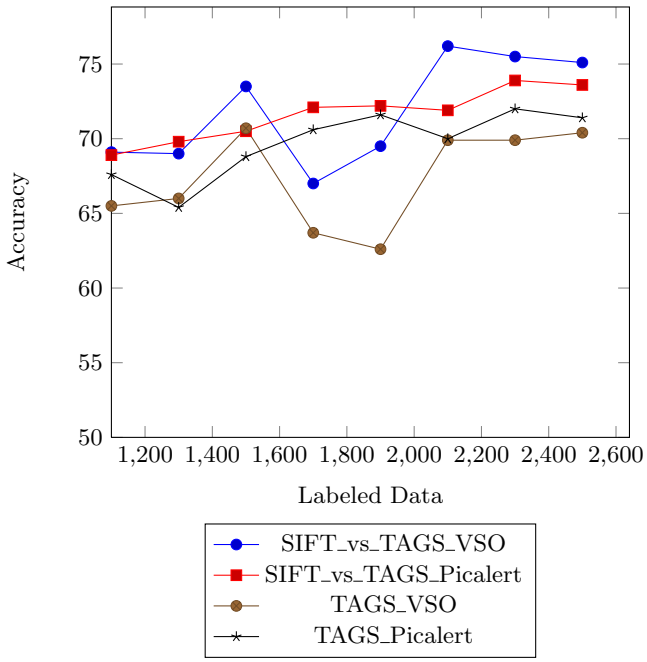

Figure 8: Comparison between performance for models on Picalert and Visual Sentiment Ontology dataset

6.1 Learning models

Our current problem can be mapped into a multi-label, multi-class problem. We have three classifications to perform, and each of them includes five possible labels, ranging from "Only You" to "Everyone". Precisely, the labels are "Only You", "Family", "Friends", "SocialNetwork" and "Everyone". Each classification is indicative of one sharing privilege, and includes "view", "comment" and "Download" access controls for each image.

Our model is based on supervised learning, and both training and test items, are represented as multi dimensional feature vectors. Labeled images, selected at random from the dataset, are used to train classification models. As mentioned, we added three categories for each image, and each category could be classified into one of the five privacy labels above. We noticed that in most of the cases the privacy levels set by users for the three categories are related. For example, if a user wants to make an image comment-able and download-able, it would be possible only if the image is view-able to the users in the same level of privacy. To account for these types of inter-relations of categories for classifying unlabeled images, we apply Chained Classifiers model [33]. Chained classifiers were developed to capture the dependency between categories in a multi-category dataset. The method enables the predictions that were made on a previous iteration on a category to be utilized in the prediction of the subsequent categories. Each classifier in the chain is a Multiclass - SVM classifier for learning and prediction.

In our context, the Chained Classifiers model [33] involves three transformations - one for each label. In a sense, chained classifier simply uses SVM on each of the labels, but it differs from multi-class SVM in that the attribute space of each label is extended by the predicted label of the previous classifiers. Given a chain of N classifiers, $\{c_1, c_2, ..., c_N\}$, each classifier c_i in the chain learns and predicts the ith label in

the attribute space, augmented by all previous label predictions c_1 to c_{i-1}. This chaining process passes information about labels between the classifiers. Although increasing the attribute space might be an overhead, if strong correlations exist between the labels then these attributes immensely help increasing the predictive accuracy. An example of a correlation in our usecase - *comment* label has the predicted value of *view* label in the attribute space. Intuitively, an image can only be commented upon if it can be viewed.

6.2 Experimental Results

We again relied on the Picalert dataset for these experiments. We sampled 4500 images, and used the same features set as in our previous experiments. For labeling purposes, we used Amazon Mechanical Turk (AMT). Quality of workers' was carefully monitored. For instance, we disregarded work from users who assigned the same set of labels for 80% of the images. We also manually checked URLs at random to check for consistency of labels.

For these set of experiments, we used the three features that performed best in our binary classification, namely SIFT, Tags, Edges. We increased the size of the training data, ranging from 2000 to 4000, keeping 500 as the size of test data.

We compared the performance of chained classifiers against a baseline classifier model. The baseline model involved running multi-class SVM on each of the three classes separately, using the same set of attributes as opposed to chained classifiers which appends the predicted class label to the attribute list for the current prediction. The two classification models (i.e. chained classifiers and baseline), were used for single-feature analysis as well as for combinations of features. Results are reported in Figure 9.

A few interesting observations can be made. First, our

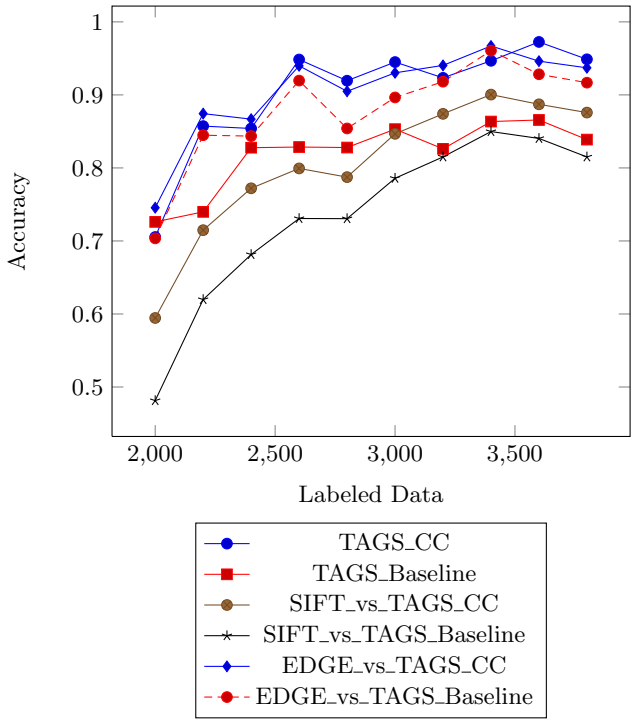

Figure 9: Multi-option privacy settings

results show the accuracy of prediction across the combinations with the increasing size of training data. Second, we noted that TAGS, as established already in our previous analysis, was the best performing single feature, with extremely high prediction accuracies (up to 94%). Ensemble of features using TAGS & SIFT and TAGS & EDGE also had high prediction accuracy, reaching up to 90%. Finally, we observed that using chained classifiers increases overall performance, regardless of the features used, in comparison to the baseline accuracy achieved using multi-class SVM on each class individually. This result confirms that chained classifiers are useful in capturing the correlation between the class labels and resulted in higher prediction accuracy.

7. CONCLUSION

In this paper, we presented an extensive study investigating privacy needs of online images. We studied the importance of images visual content and user-applied metadata to identify adequate privacy settings. We considered both the case of privacy settings being simple, as well as the case of more complex privacy options, which users may choose from. Our results show that with few carefully selected features, one may achieve extremely high accuracy, especially with the support of metadata.

In the future, we will extend this work along several dimensions. First, given the strong performance of TAGS, we would like to incorporate additional textual metadata in the models, to further the performance of this class of features. Second, we would like to further explore the visual models and their roles in the context of complex privacy settings. Finally, we will implement a recommender system for users to express privacy settings based on their privacy choices.

8. REFERENCES

[1] Java content based image retrieval, 2011. https://code.google.com/p/jcbir/.

[2] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 357–366, New York, NY, USA, 2007. ACM.

[3] E. A. Alessandra Mazzia, Kristen LeFevre, April 2011. UM Tech Report #CSE-TR-570-11.

[4] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 971–980, 2007.

[5] A. Besmer and H. Lipford. Tagged photos: concerns, perceptions, and protections. In *CHI '09: 27th international conference extended abstracts on Human factors in computing systems*, pages 4585–4590, New York, NY, USA, 2009. ACM.

[6] S. D. Blog. Pin or not to pin: An inside look, 2012. http://blog.socialdiscovery.org/tag/statistics/.

[7] J. Bonneau, J. Anderson, and L. Church. Privacy suites: shared privacy for social networks. In *Symposium on Usable Privacy and Security*, 2009.

[8] J. Bonneau, J. Anderson, and G. Danezis. Prying data out of a social network. In *ASONAM: International*

Conference on Advances in Social Network Analysis and Mining, pages 249–254, 2009.

[9] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs, 2013. `http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html`.

[10] Bullguard. Privacy violations, the dark side of social media. http://www.bullguard.com/bullguard-security-center/internet-security/social-media-dangers/privacy-violations-in-social-media.aspx.

[11] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, 1999.

[12] S. Chatzichristofis, Y. Boutalis, and M. Lux. Img(rummager): An interactive content based image retrieval system. In *Similarity Search and Applications, 2009. SISAP '09. Second International Workshop on*, pages 151 –153, aug. 2009.

[13] G. P. Cheek and M. Shehab. Policy-by-example for online social networks. In *17th ACM Symposium on Access Control Models and Technologies, SACMAT '12*, pages 23–32, New York, NY, USA, 2012. ACM.

[14] H.-M. Chen, M.-H. Chang, P.-C. Chang, M.-C. Tien, W. H. Hsu, and J.-L. Wu. Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 737–740, New York, NY, USA, 2008. ACM.

[15] M. D. Choudhury, H. Sundaram, Y.-R. Lin, A. John, and D. D. Seligmann. Connecting content to community in social media via image content, user tags and user communication. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, pages 1238–1241. IEEE, 2009.

[16] R. da Silva Torres and A. Falcão. Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada*, 2(13):161–185, 2006.

[17] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[18] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 71–84, Berlin, Heidelberg, 2010. Springer-Verlag.

[19] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 351–360, New York, NY, USA, 2010. ACM.

[20] J. He, W. W. Chu, and Z. Liu. Inferring privacy information from social networks. In *IEEE International Conference on Intelligence and Security Informatics*, 2006.

[21] X. He, W. Ma, O. King, M. Li, and H. Zhang. Learning and inferring a semantic space from user's relevance feedback for image retrieval. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 343–346. ACM, 2002.

[22] K. J. Higgins. Social networks for patients stir privacy, security worries, 2010. Online at `http://www.darkreading.com/authentication/167901072/security/privacy/227500908/social-networks-for-patients-stir-privacy-security-worries.html`.

[23] S. Jones and E. O'Neill. Contextual dynamics of group-based sharing decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1777–1786. ACM, 2011.

[24] P. F. Klemperer, Y. Liang, M. L. Mazurek, M. Sleeper, B. Ur, L. Bauer, L. F. Cranor, N. Gupta, and M. K. Reiter. Tag, you can see it! Using tags for access control in photo sharing. In *CHI 2012: Conference on Human Factors in Computing Systems*. ACM, May 2012.

[25] K. Liu and E. Terzi. A framework for computing the privacy scores of users in online social networks. *ACM Trans. Knowl. Discov. Data*, 5:6:1–6:30, December 2010.

[26] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[28] A. D. Miller and W. K. Edwards. Give and take: a study of consumer photo-sharing culture and practice. In *CHI '07: SIGCHI conference on Human factors in computing systems*, pages 347–356, New York, NY, USA, 2007. ACM.

[29] W. W. Ng, A. Dorado, D. S. Yeung, W. Pedrycz, and E. Izquierdo. Image classification with the use of radial basis function neural networks and the minimization of the localized generalization error. *Pattern Recognition*, 40(1):19 – 32, 2007.

[30] A. Plangprasopchok and K. Lerman. Exploiting social annotation for automatic resource discovery. *CoRR*, abs/0704.1675, 2007.

[31] M. Rabbath, P. Sandhaus, and S. Boll. Automatic creation of photo books from stories in social media. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7S(1):27:1–27:18, Nov. 2011.

[32] M. Rabbath, P. Sandhaus, and S. Boll. Analysing facebook features to support event detection for photo-based facebook applications. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 11:1–11:8, New York, NY, USA, 2012. ACM.

[33] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification, 2011.

[34] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 771–780, New York, NY, USA, 2009. ACM.

[35] N. Sawant. Modeling tagged photos for automatic image annotation. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 865–866, New York, NY, USA, 2011. ACM.

[36] N. Sawant, J. Li, and J. Z. Wang. Automatic image semantic interpretation using social action and tagging data. *Multimedia Tools Appl.*, 51(1):213–246, 2011.

[37] J. Sivic and A. Zisserman. Video google: A text

retrieval approach to object matching in videos. In *Proc. of ICCV*, pages 1470–1477, 2003.

[38] A. C. Squicciarini, S. Sundareswaran, D. Lin, and J. Wede. A3P: adaptive policy prediction for shared images over popular content sharing sites. In *22nd ACM Conference on Hypertext and Hypermedia*, pages 261–270. ACM, 2011.

[39] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computational visual attention model. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 541–544, New York, NY, USA, 2009. ACM.

[40] H. Sundaram, L. Xie, M. De Choudhury, Y. Lin, and A. Natsev. Multimedia semantics: Interactions between content and community. *Proceedings of the IEEE*, 100(9):2737–2758, 2012.

[41] A. Vailaya, A. Jain, and H. J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31(12):1921 – 1935, 1998.

[42] N. Vyas, A. C. Squicciarini, C.-C. Chang, and D. Yao. Towards automatic privacy management in web 2.0 with semantic analysis on annotations. In *CollaborateCom*, pages 1–10, 2009.

[43] C. Wang, D. M. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1903–1910. IEEE, 2009.

[44] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proc. of ACM Workshop on Multimedia Information Retrieval*, pages 197–206, 2007.

[45] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *Proceedings of the international conference on Multimedia*, MM '10, pages 211–220, New York, NY, USA, 2010. ACM.

[46] C. Yeung, L. Kagal, N. Gibbins, and N. Shadbolt. Providing access control to online photo albums based on tags and linked data. *Social Semantic Web: Where Web*, 2, 2009.

[47] J. Yu, X. Jin, J. Han, and J. Luo. Social group suggestion from user image collections. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1215–1216, New York, NY, USA, 2010. ACM.

[48] J. Yu, D. Joshi, and J. Luo. Connecting people in photo-sharing sites by photo content and user annotations. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1464–1467. IEEE, 2009.

[49] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova. Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 35–44, New York, NY, USA, 2012. ACM.

[50] N. Zheng, Q. Li, S. Liao, and L. Zhang. Which photo groups should I choose? a comparative study of recommendation algorithms in flickr. *J. Inf. Sci.*, 36:733–750, December 2010.

[51] J. Zhuang and S. C. H. Hoi. Non-parametric kernel ranking approach for social image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 26–33, New York, NY, USA, 2010. ACM.