Toward Automated Online Photo Privacy

ANNA SQUICCIARINI, Pennsylvania State University
CORNELIA CARAGEA, University of North Texas
RAHUL BALAKAVI, Pennsylvania State University

Online photo sharing is an increasingly popular activity for Internet users. More and more users are now constantly sharing their images in various social media, from social networking sites to online communities, blogs, and content sharing sites. In this article, 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. Our study investigates both visual and textual features of images for privacy classification. We consider both basic image-specific features, commonly used for image processing, as well as more sophisticated and abstract visual features. Additionally, we include a visual representation of the sentiment evoked by images. To our knowledge, sentiment has never been used in the context of image classification for privacy purposes. 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 high accuracy, especially when high-quality tags are available.

CCS Concepts: • **Security and privacy → Software and application security**; **Social network security and privacy**

Additional Key Words and Phrases: Social networks, image analysis, privacy, machine learning

ACM Reference Format:
Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2017. Toward automated online photo privacy. ACM Trans. Web 11, 1, Article 2 (March 2017), 29 pages.
DOI: http://dx.doi.org/10.1145/2983644

1. INTRODUCTION

Online photo sharing is an increasingly popular activity for Internet users. More and more users are now constantly sharing their images in various social media, from social networking sites to online communities, blogs, and content sharing sites. 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 [Blog 2012] 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

This work is supported by the National Science Foundation, under grant 1421776 and grant 1421970.
Authors' addresses: A. Squicciarini, 301D IST Building, College of Information Sciences and Technology, Pennsylvania State University, University Park PA 16802; email: asquicciarini@ist.psu.edu; C. Caragea, F228 Discovery Park, Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA; email: ccaragea@unt.edu; R. Balakavi, 317 IST Building, College of Information Sciences and Technology, Pennsylvania State University, University Park PA 16802; email: rahulbalakavi@gmail.com.

DOI: http://dx.doi.org/10.1145/2983644

[Blog 2012; Zheng et al. 2010]. However, semantically rich images may reveal content-sensitive information [Ahern et al. 2007; Squicciarini et al. 2011; Zerr et al. 2012]. Consider a photo of a student's 2014 New Years' public ceremony, for example. It could be shared within a Google+ circle or Flickr group, or it could be used to discover 2014 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 [Bullguard 2014; Ahern et al. 2007; Besmer and Lipford 2009]. Malicious attackers can take advantage of these unnecessary leaks to launch context-aware attacks or even impersonation attacks [Higgins 2010], as demonstrated by a proliferating number of cases of privacy abuse and unwarranted access.

In particular, privacy of online images is inherently a subjective matter, dependent on the image owner's privacy attitude, awareness, and the overall context wherein the image is to be posted.

In this work, we explore the hypothesis that some generic patterns of private images can be well identified when a group of online images are taken into consideration, regardless of their authors' individual privacy bias and level of awareness.

Toward validating this hypothesis, 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. We focus on image-specific features only, rather than broader contextual social network dimensions or personal information about the image poster or his/her audience. Intuitively, contextual features may help in addressing our research question but would require much more information for every image, which may or may not always be available or even reliable. We aim to minimize additional personal information that would be needed to infer users' privacy preferences. To achieve this goal, we focus on two types of image features: visual-content features and images' 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. Among the features that we use to capture the visual-content of images, we include both low-level image processing features that capture colors, patterns, and edge directions, with more sophisticated derived features. One such derived feature is "sentiment," representing the sentiment evoked by the image. Our hypothesis here is that the sentiment evoked by an image may be correlated with the type of disclosure associated with it. To the best of our knowledge, sentiment has never been studied in correlation with privacy classification of images.

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 significant discriminative power.

To this date, only very few studies have started to address this complex problem. Most recent work related to online disclosure of personal information has been devoted to protecting generic textual users' online personal data, with no emphasis on the unique privacy challenges associated with image sharing [Liu and Terzi 2010; He et al. 2006]. 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., Chapelle et al. [1999], Sawant et al. [2011], Ng et al. [2007], Datta et al. [2008], and da Silva Torres and Falcão [2006]). Only some recent work has started to explore simple classification models of image privacy [Squicciarini et al. 2011; Zerr et al. 2012].

We specifically identified Scale-Invariant Feature Transformation (SIFT) 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. In the absence of TAGS, our results show that SIFT and Sentiment-based feature perform the best, with prediction accuracy reaching unto 80%.

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, our models also include various 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 [Read et al. 2011] models, where predicted class labels are used to predict new class labels. Our experiments confirm that these models, executed using a blend of visual and metadata features, 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 article 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 article.

2. RELATED WORK

A number of recent studies have analyzed sharing patterns and social discovery in image sharing sites like Flickr [Choudhury et al. 2009; Ames and Naaman 2007; Miller and Edwards 2007; Zheng et al. 2010]. 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 [Sawant 2011; Plangprasopchok and Lerman 2007; Chen et al. 2008; Ames and Naaman 2007; Henne et al. 2013], motivating our hypothesis of using metadata as one among other features for privacy extraction.

Miller and Edwards [2007] further confirm that people who share their photos maintain social bonds through tagging together with online messaging, commenting, and so on. They also identify two different types of users (normal and power users), indicating the importance of interpersonal differences, and that users may have different levels of privacy concerns depending on their individual level of privacy awareness and the image content.

Ahern et al. [2007] 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 O'Neill [2011] 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, and 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 [Zerr et al. 2012], which carries out content analysis for image private search and detection of private images.

Along the same theme, Besmer and Lipford [2009] 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 [Alessandra Mazzia 2011; Vyas et al. 2009; Cheek and Shehab 2012; Squicciarini et al. 2011; He et al. 2006; Liu and Terzi 2010]. Some notable recent efforts to tackle these problems have been conducted in the context of tag-based access control policies for images [Yeung et al. 2009; Klemperer et al. 2012; Vyas et al. 2009], showing some initial success in tying tags with access control rules. However, the scarcity of tags for many online images [Sundaram et al. 2012], and the workload associated with user-defined tags precludes accurate analysis of the images' sensitivity based on this dimension only. Other work [Fang and LeFevre 2010; Cheek and Shehab 2012; Alessandra Mazzia 2011; Bonneau et al. 2009a, 2009b; He et al. 2006] has focused on generic users' profile elements and typically leveraged social context rather then users' individual content-specific patterns.

A loosely related body of work is on recommendation of tags for social discovery [Sawant 2011; Plangprasopchok and Lerman 2007; Yu et al. 2010; Chen et al. 2008] and for image classification [San Pedro and Siersdorfer 2009; Yu et al. 2009; Chen et al. 2008] in photo sharing websites like Flickr. In these works, the typical approach is for authors to first collect adequate images and then classify images according to visual and context-related features. After users upload their images, the server extracts features and then classifies and recommends relevant tags and groups.

Also related is the work from Henne and colleagues [Henne et al. 2013], who provided an extended analysis of privacy threats as they arise from photos in popular online content sharing sites, such as Flickr. As noted by Henne and others [Madejski et al. 2012; Xu et al. 2015], privacy threats result from either a user's own actions or shared photos or are unintentional, as a consequence of others' photo uploads. We note that both types of threats are important and of increasing relevance and can yield to "errors" or unintended issues. In this work, we keep our focus on image protection based on access policies, under the implicit assumptions that these are applied by one authorized entity, that is, the image uploader. We note, however, that our work is agnostic to the problem of ownership: A policy could be applied either by the party who owns the image or by a third party seeing him- or herself tagged or exposed in an image uploaded by others. We acknowledge that this may raise some interesting new research questions that we plan to explore in the future.

Finally, there is a large body of work on image content analysis for classification and interpretation (e.g., Chapelle et al. [1999], Sawant et al. [2011], Ng et al. [2007], Vailaya et al. [1998], Zhuang and Hoi [2010], Wang et al. [2009], and Deng et al. [2010]), retrieval (Datta et al. [2008], da Silva Torres and Falcão [2006], He et al. [2002], and Chatzichristofis et al. [2009] are just some examples), and photo ranking [Sun et al. 2009; Yeh et al. 2010], also in the context of online photo sharing sites, such as Flickr [San Pedro and Siersdorfer 2009; Yu et al. 2009; Chen et al. 2008; Sundaram et al.

2012; Rabbath et al. 2011, 2012; Choudhury et al. 2009]. This previous work is useful in identifying meaningful content-based features for effective image content extraction, discussed in Section 4.

We presented a preliminary version of this work in Squicciarini et al. [2014]. In this extended work, we extend previous contribution in a number of ways. First, we introduce a new perspective in the analysis of the images, in that we introduce a new complex semantic feature that has never been linked to image privacy, that is, sentiment feature. We show how the sentiment an image evokes helps in linking an image's disclosure setting. Further, we also perform an in depth analysis of outliers, carry a large amount of new experiments, and provide a new in-depth set of considerations on the role of the analyzed features in helping uncover privacy patterns.

3. PROBLEM STATEMENT

The objective of our work is to infer adequate privacy settings for online images, based on a community perspective of general users' privacy preferences.

Our goal is twofold: (1) We aim to identify a variety of visual features that can be informative in profiling images' privacy needs, and (2) among the identified features, we 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. We note that these are challenging objectives, as the classification we hope to achieve is based on the subjective notion of privacy that therefore attempts to assign a "semantic" meaning to an image rather than simply describe its main content (or extract the context).

Our approach is to consider both images' visual content and their associated metadata. The intuition underlying content-based features is that, as demonstrated in recent work (e.g., Besmer and Lipford [2009]), 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. Our content-specific features include a selection of both "basic" features commonly adopted in image processing, as well as more abstract, sophisticated features. Among others, we experiment with sentiment features. Sentiment-related features refer to the ability of capturing the emotions and feelings reflected by an image.

On the contrary, metadata, typically defined in terms of keywords extracted from tags or captions, can provide insights into the image's context, that is, where it was taken, what it represents to the labeler (e.g. the image owner), what feelings it evokes, and so on.

Additional contextual dimensions are purposely not considered for the purpose of this study. For instance, we do not consider any additional social networking or personal information about the photo owners and the site where the image was 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 or her 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 this accounts for the case of an image that 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 an online social networking site, where users may choose from a fine-grained set of options (i.e., should a 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 a few pivotal features that can be useful for image classification with respect to privacy. We next describe our selected visual features and provide some observations from the use of the features for privacy models.

—*SIFT* [Lowe 2004b]. As an image's privacy level may be determined by the presence of one or more specific objects rather than all of the visual content (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 the scale space of an image. In simple terms, the SIFT approach is to take an image and transform it into a "large collection of local feature vectors" [Lowe 2004a]. Each of these feature vectors is invariant to any scaling, rotation, or translation of the image. We extract an image profile based on the state-of-the-art model called bag-of-visual-words (BoW) [Sivic and Zisserman 2003; Yang et al. 2007], which can effectively evaluate the image similarity and is widely used in image similarity search, image retrieval, and even content-based social discovery [Sawant 2011]. The image BoW vector is first obtained by extracting the features of preferred images and 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 BoW. 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 $\triangle$.

—*Sentiment.* The emotions evoked by an image may be tightly related to an image's privacy needs. Accordingly, we leverage a Visual Sentiment Ontology (VSO) to detect some sentiments. A VSO is constructed based on understanding that the visual concepts are strongly related to sentiment [Borth et al. 2013]. More precisely, VSO is built on psychological theories and web mining comprising sentiment carrying concepts called Adjective Noun Pairs (ANPs). *Cute Dog, Beautiful Day, Disgusting Food* are a few examples of ANPs. A set of such strong sentiment-carrying concepts are collected and each of these concepts are encoded with visual information pertaining to images that are relevant to that ANP. The detector library identifies a total of 1,200 concepts. Given a test image, the ANP detector framework outputs a series of 1,200 decimal numbers, each in the range of 0–1. Each of these numbers indicates the degree of presence of the corresponding ANP in the test image.

—*Red Green Blue* (RGB). Images with a given color and texture patterns can be mapped into certain classes, based on what is learned 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 the RGB feature to extract these potentially useful patterns. RGB is a color space for image representation, and Hue, Saturation, and Brightness (HSB) values can be obtained, for example, by a color space conversion from RGB to HSB, obtaining a feature detector. The feature detector components are therefore Red, Green, Blue, Hue, Saturation and Brightness. Values corresponding to each of the variables are extracted from an image, and each

Fig. 1. FACIAL detection and extraction -examples.

feature is encoded into a 256byte length array. This array is serialized in sparse format where each instance corresponds to the feature vector of an image.

—*FACIAL Detection.* Images revealing one's identity are more likely to be considered private [Ahern et al. 2007], 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, similarly to Zerr et al. [2012]. 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 that are straight up and clearly visible. In some images, though there are faces visible, due to various factors, the faces could not be detected. Images with faces not clearly visible, images with dark backgrounds, and images that show faces from acute angles are a few factors that can result in faces not being detected by the API (Application Programming Interface). In Figure 1, we show two sets of images where FACIAL detection is successful and where it fails.

—*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 proven to be 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

Fig. 2. Example of a Flickr image and its tags.

pixels, we backtrack from a random coherent pixel to another and check if it is within 5° and then update the corresponding coherent bins [Vailaya et al. 1998].

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, as well as a set of Flickr-generated tags. In this work, we focused on user-generated tags only and created a feature vector of tags for each image accordingly. We created a dictionary of words from all images in the training set such that there are no duplicates (in the dictionary) and applied basic stemming methods to limit the noise introduced by misspelling, use of plurals, and so on. 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 for a compact feature vector for each instance, removing unnecessary information about the 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 that show similar descriptive patterns have extensive word usage that is 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 to be extremely useful: As we further elaborate in Section 5.2.4, tags are a predominant feature for privacy classification purposes, although acceptable results are found even in the 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 the first dataset, we took a sample of Flickr images from the PicAlert dataset [Zerr et al. 2012]. 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 6,000 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 [Borth et al. 2013]. The repository has a good blend of landscape images, animals, artwork, people and so on. About 4,000 Flickr URL (Uniform Resource Locator) 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 (AMT)) for labeling the additional datasets and complement existing labels with more complex settings (see Section 6).

Of course, similarly to Zerr et al. [2012], the adopted privacy labels only capture an aggregated community perception of privacy rather than 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. We also believe that using a dataset reflecting a community, rather than a dataset created from a single author, poses some interesting unique challenges. Because images are not linked to individual users and the image authors do not label them with their privacy preferences, it is non-trivial to find classification models and the set of features that can help extract them.

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.

Our approach to identify these characteristics is to group images by content similarity to explore the visual similarities that define the clusters. To this end, we used unsupervised learning methods. In particular, we applied the Java API for Content Based Image Retrieval (CBIR) from Latha and colleagues [Latha 2011]. This implementation performs image clustering to identify clusters among public and private images, respectively. The API 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. An image retrieval algorithm is used for image retrieval based on similarity.

On running CBIR, we observed similarity patterns among images in different sets that were clustered. Most public images belong to one of three categories as follows: (1) Women and Children, (2) Symbols and black-and-white images, (3) Artwork. Private images could be mainly grouped into (1) People and (2) Sketches. Note that while there is some overlap among these categories, the images with women and children in the public categories portray mostly photos taken in public settings, whereas private images portraying people are close-up images with more skin exposed.

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, and so on, 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. Finally, sketches appear to be deemed sensitive. We speculate these may be labeled as private, as they may represent personal aspects of one's life, like a tattoo or a piece of art.[1] 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 [Squicciarini et al. 2011], using Wordnet as a 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 stored 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 that appeared most frequently. In case of ties, we choose the word that is closest to the root or baseline.

We repeated the same procedure over different parts of speech, that is, 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 the 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, for example, "kid." In the case where there are more than one hypernym 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, 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,

[1]Note that our observations are mainly qualitative; to fully grasp the difference of these classes, one may need human analysis and additional image processing work, which goes beyond the scope of this work.

Table I. Common Keywords in Private Images

Cluster	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

Fig. 3. Example images belonging to the first and second clusters of private images, respectively (relation, water, and appearance).

and adjective. These parts of speech are chosen as they are the words with highest frequency among the instances of the cluster.

Using the above methodology, we clustered about 6,000 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 I 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). Figure 3 shows sample images being tagged with keywords from the first and second clusters, respectively.

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 (examples are reported in Figure 3). 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 II, 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 that described dressing style or appearance patterns. Cluster 4 described art work or objects with patterns. Examples of images

Table II. Common 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

Fig. 4. Example images belonging to the first and second clusters of public images (season and electricity, respectively).

linked to the keywords in public images are shown in Figure 4. 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. Our datasets used for testing and training always preserve a 2:1 ratio, unless specified otherwise. That is, for every private (positive) image, there are two public (negative) images. We maintain this ratio, as it is usually the case that most images are actually public, and only a fraction of them contains private content.

5.2.1. Individual Features Analysis. We begin by studying the performance of individual features. We specifically evaluated the performance of classifiers trained on 4,000 labeled images and evaluated them on a test set of 500 images, using one feature per model.

Table III. Overall Accuracy of Various Learning Models

	Feature		
Learning Model	SIFT	TAGS	FACIAL Features
Naïve Bayes	49.8	63.9	60.6
Naïve Bayes Multinomial	55.5	76.04	53.5
Logistic Regression	59.6	74	60.3
SVM	59.8	81	62.75
k-nearest neighbors	59	73.75	59.7

Table IV. Individual Features' Classification Models over 4,500 Images

Features	Accuracy	Precision	Recall
TAGS	78.2	0.791	0.782
RGB	56.6	0.576	0.548
SIFT	59	0.613	0.59
FACIAL Feat.	61.2	0.584	0.612
EDGE Dir. Coh.	54.2	0.588	0.542
Sentiment	73.6	0.72	0.701

There are a variety of supervised learning models such as linear regression, näive Bayes, decision trees, and so on. Each model performs well for a particular scenario and conditions. Factors like heterogeneity of data, data redundancy, presence of interactions among features, and so on, should be considered when choosing and applying a learning algorithm. Since the features that are produced using image extraction are independent and discrete, most of the mentioned models can be helpful for classification purposes. We use the standard k-fold cross-validation technique (with k set to 10) to estimate performance of a learning algorithm in the dataset. Both training and test items are represented as multi-dimensional feature vectors. Labeled images are used to train classification models. Table III shows k-fold cross-validation results on different supervised learning models. These results are acquired by studying the performance of individual features on different supervised learning models. We observe that Support Vector Machine (SVM) performs consistently better than the other models in predicting image privacy, regardless of the specific feature being considered. Accordingly, we determined that SVM is a strong fit for our supervised learning analysis.

SVMs are a class of supervised learning models that analyze data and recognize patterns used for classification. 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. The main aim of the model is to separate training data with minimum or no errors.

Using SVM (with RBF kernel) as a learning model, we then analyze more in depth the performance of individual features. Results are reported in Table IV. As shown, TAGS is by far the best-performing feature. Content-based features have lower accuracy, with the worst observed for the EDGE Direction feature, for which accuracy is only 54.2%. Similar results are observed when we vary the size of the training data for various features, as reported in Figure 5. As shown, the best features are consistently TAGS and SIFT features, regardless of the size of the training data.

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

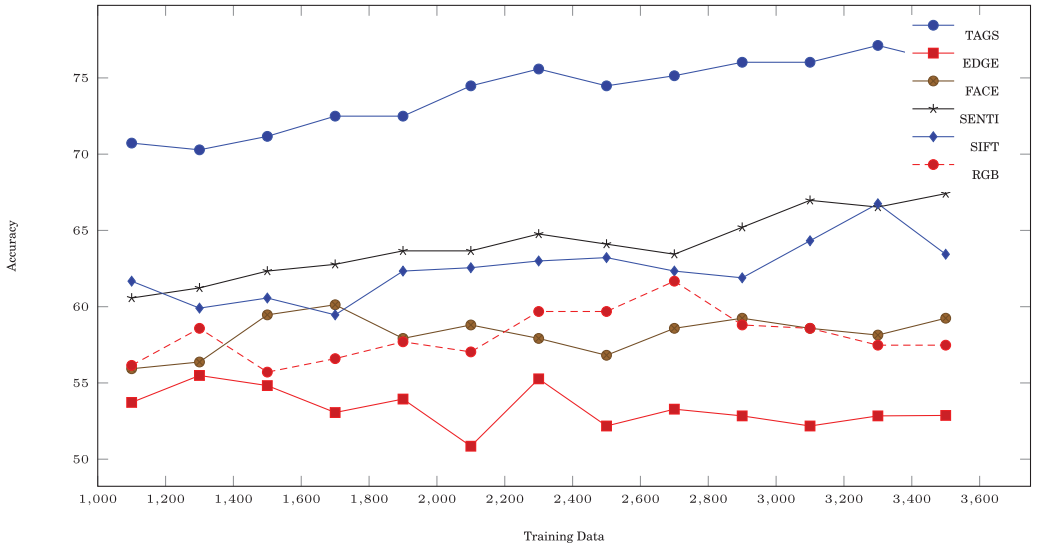

Fig. 5. Accuracy analysis of linear models with single features.

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. Image Pruning for Outlier Detection. Our dataset is very diverse, and it may include many noisy images and misplaced labels, especially given the original labeling method adopted (i.e., crowdsourcing). In order to test whether outliers may be successfully removed from the image set, we deployed a distance-based pruning method for outliers (from the training set).

Our outlier detecting mechanism, inspired by Ramaswamy et al. [2000], suggests that anything that seems to go out of pattern in a clustering process is called a deviation. Our task is to find out such deviations or outliers. As per the generic definition of outliers, for a given point for a particular distance d and a number of points k, a data point is called an outlier if it has no more than k points within the distance radius d from it. Generally a user has to determine the values of d and k via trial and error. We implement the method as described by Ramaswamy et al. [2000]. The value of the distance d need not be hard coded and tested every time. The value of k is varied and the number of outliers n is instead already established. If $D(p)$ is the distance of the kth nearest neighbor of p, then that is the distance to be considered. So for a given value of n, where p is the point and $D(p)$ the distance from the kth neighbor, it is considered an outlier if no more than $n - 1$ other points p' have a distance greater than $D(p')$. In our experiment, the training set is sorted as per the distances $D(p)$, and the value of n decides the outliers to be pruned.

We also compared this method with a baseline, an average pruning method, that instead prunes images from the training set if they are far enough from the average value of the image for any given image. We report some of the findings in Table V. The

Table V. Outliers Detection Using Pruning- Accuracy Is Computed by Averaging the Class Results

No of classes	Features	Dataset size	Accuracy
4	SIFT+ averaging	1,600	0.58
4	SIFT+knnprune	1,600	0.63
2	dSIFT+average	4,800	0.64
2	SIFT	4,800	0.67
2	SIFT + Knnprune	4,800	0.69

results summarize the difference in performance for a simple classification task based on SIFT only (since it is one of the most successful features). In the experiments, we varied both the number of classes (private, public or clear public, public, private, clear private) and the size of the dataset. To create an unbiased dataset sample, we use a ratio of 1:1 for every class.

As can be seen, pruning brings a significant improvement of almost 9% for the case of two class labelings. This early evidence seems to justify a systematic pruning method to remove unnecessary noisy images.

5.2.3. Models Combination. We explored various combinations of supervised learning models for image privacy, using the features listed in Table IV. For these experiments, we did not do any specific image pruning. 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 1,000 to 3,700, keeping 500 as the size of the test data. To combine different features, we linearly combined the vectors of feature representations of individual features into a single vector. For example, given $F_1 = \{s_0, s_1, \ldots, s_j\}$ and $F_2 = \{r_1, r_1, \ldots, r_k\}$, where F_1 and F_2 are two different feature representations of lengths j and k for an image, combining them linearly results in a new feature vector $F_{com} = \{s_0, s_1, \ldots, s_j, r_1, r_1, \ldots, r_k\}$ of length $j + k$.

Our results, reported in Figure 6 for two-feature 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.

In Figures 7 and 8, 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 6), where the best accuracy is obtained with SIFT+TAGS on a labeled dataset of 3,300 training instances, followed by EDGE+TAGS.
—SIFT+TAGS appears to be strongest combination for both two-feature and three-feature models. Intuitively, SIFT captured all the visual key points of the images,

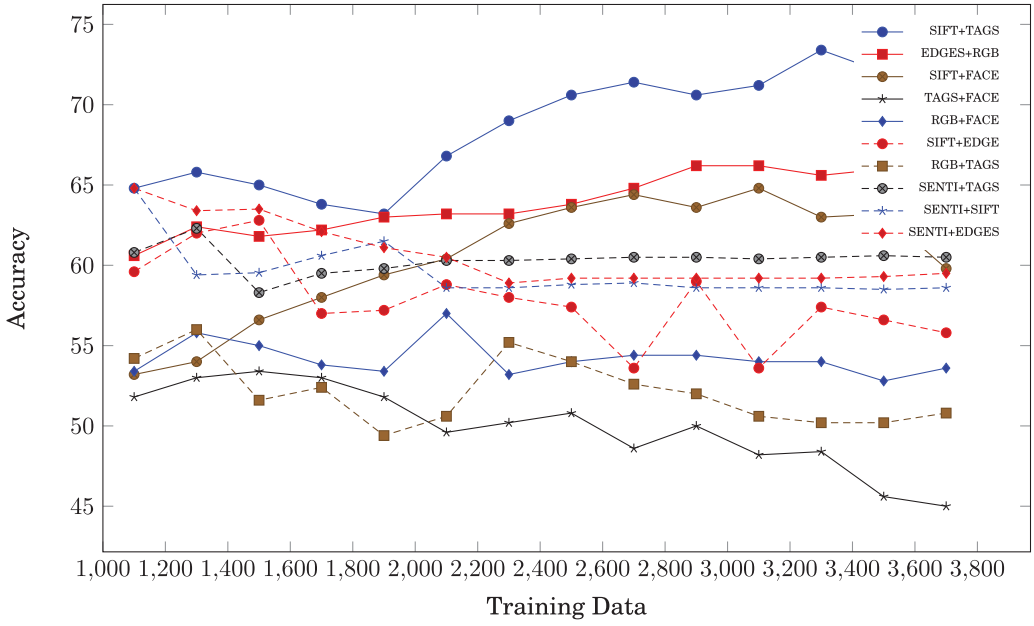

Fig. 6. Feature analysis of models with two features concatenated.

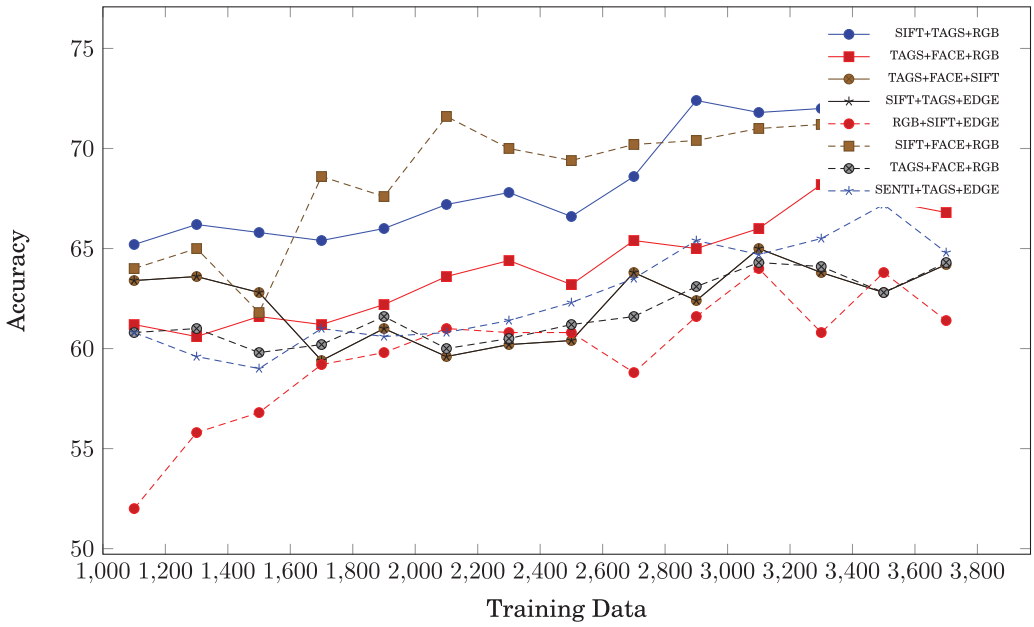

Fig. 7. Accuracy analysis of combined classifiers with three features.

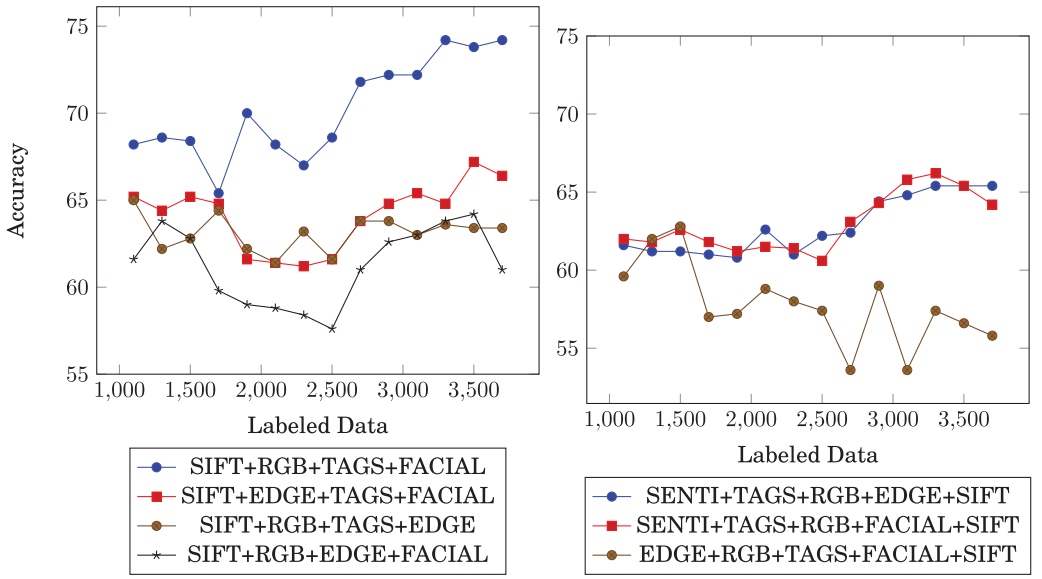

Fig. 8. Accuracy analysis of linear models with five features.

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 70% over a dataset of size 4,500 for SIFT, EDGE, and RGB combined together. The combined classifier resulted in a BEP of 0.667.

—Model combinations including sentiment seem not to provide exciting results compared to other (sentiment-free) combinations. For instance, as reported in Figure 7, no combination including sentiment feature achieves the performance of TGS, FACE, and RGB combined together. The best results for sentiment features combination are about 60–65% accuracy against about 75% accuracy achieved by the TAGS, FACE, and RGB combination.

—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 8 for the performance of all features combination). For instance, SIFT+FACE+RGB performs worse oversall than their individual features.

5.2.4. 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 that is computed in consideration of both learning models.

In particular, in our case, an ensemble of classifiers is a collection of classifiers, each trained on a different feature type, for example, 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 $y_i \in \{private, public\}$,

Table VI. Performance of Ensemble Models Combining TAGS with One Content Feature

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
TAGS and Senti	84.6	0.844	0.846

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 that 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 of normalized feature vectors: 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 4,500 images (4,000 training and 500 tests) reported in Table VI confirm this intuition. In particular, we 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 5.2.3). We note that (although not reported in detail) other ensemble classifiers that did not include TAGS do not reach interesting performance. We speculate that this is due to the very nature of the features, which fail to complement one another in the ensemble model.

In addition, to assess the validity of our ensemble, we also compare our logistic model with other types of ensembles using a training dataset ranging from 1,000 to 2100 images. We use a testing set of 500 items and again perform 10-fold cross validation. The results are shown in Figure 9. Precisely, we tested our best-performing combination of concatenated models, that is, SIFT+TAGS with two alternative methods. The first method is boosting. With boosting, we aimed at reducing bias in the data and improving the weak learning models computed with SIFT and TAG by generating one strong classification model. Boosting calculates the outputs of the TAGS and SIFT models and then averages the result using a weighted average approach. By combining the advantages and pitfalls of these approaches, we should obtain a good predictive power for a wider range of input data. Second, we performed stacking, which should help even further, as it calculates the individual models and then uses a single-layer logistic regression model as the combiner to compute the final label. As shown, the overall accuracy is lower than the accuracy reported in Table VI, due to the smaller dataset being used. Nevertheless, the accuracy of our ensemble model is consistently stronger than the other methods (reaching 76%), only capped by stacking in a couple of rounds.[2]

In Figure 10, we show the precision vs. recall graph for the ensemble classifier of SIFT and TAGS. Accordingly, the BEP, the point where the value of precision is equal to that

[2]Further experiments, which are outside the scope of this work, would help to clarify the reasons beyond differences in performance among these models.

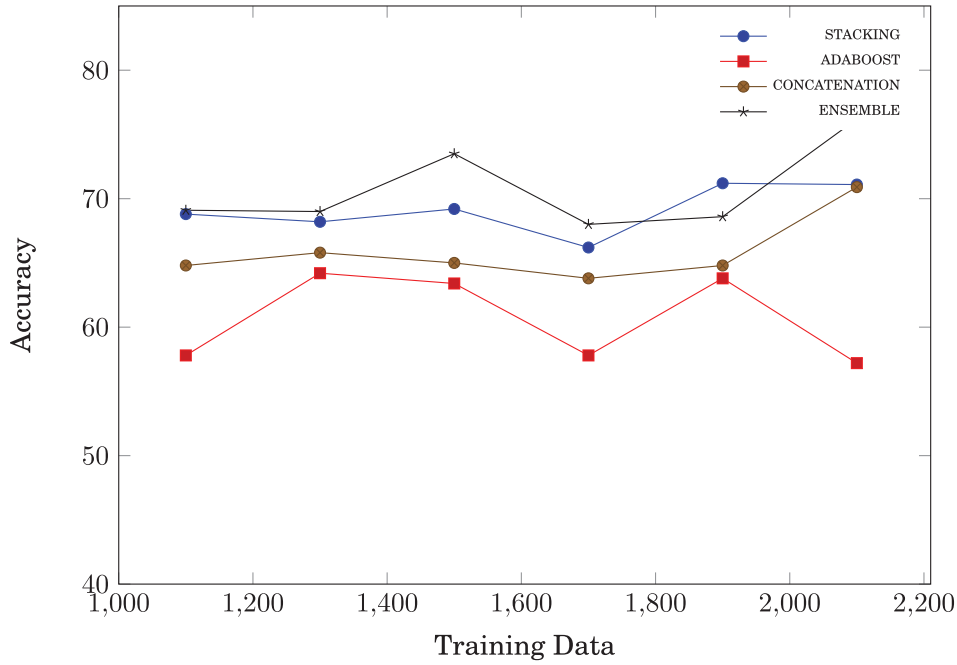

Fig. 9. Comparison of ensemble with adaboost and stacking.

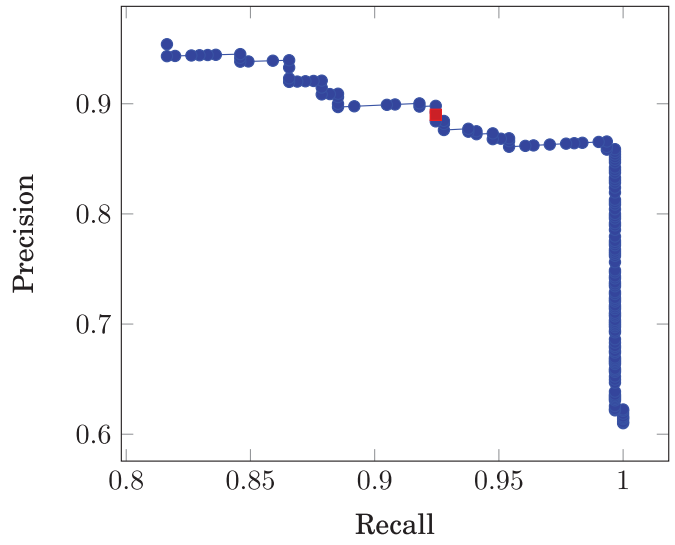

Fig. 10. Precision vs. Recall for SIFT+TAGS ensemble models.

of recall, stands at 0.89. Compared to the PicAlert framework [Zerr et al. 2012], which presented a BEP of 0.80 after combining 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

Fig. 11. False-positive and false-negative images of TAGS.

FACIAL features, which achieve very low precision and recall, showing that the choice of visual features is to be carefully made.

5.2.5. Error Analysis. We further analyzed our results, as obtained by our strongest learning model, which is an ensemble of SIFT and TAGS, and compared these results with the performance of the classifier labeled using just TAGS, which by and large was the best single-feature classifier in terms of accuracy. Further, we used unsupervised learning (CBIR) [Latha 2011] to discover emerging patterns in the labeling errors introduced by these models and understand the factors that influence an image to be either public or private.

5.3. Clustering of Images

Figure 11 shows the results of clustering images that were wrongly classified using TAGS. We separated the wrongly classified images into two sets as follows: Public images classified as Private (false positive) and Private images classified as Public (false negative). The image reports false positives on the top and false negatives at the bottom. False positives were mainly 1. artwork, 2. women, and 3. non-living things or objects. False negatives were mainly 1. images filled with alot of color and 2. images with dark background. We also observe that the majority of the images that were mis-classified were the ones that have few tags or if the user tagged the images with unconventional words, like using emoticons or a different language. These few cases are not very helpful and might result in misclassification.

Based on these error patterns, we followed up by analyzing our model based on an ensemble of SIFT+TAGS and see the improvement it brings. Recall that, as shown in Table VI, the two classifiers, SIFT+TAGS and only TAGS, have a gap in accuracy of almost 8%. As mentioned, as tags may not always be present in public images, we aim to identify a classifier that would perform well even when the keywords associated with images are not useful. The keywords might not be useful for many reasons, such as the usage of emoticons in words, which would pose problems in mapping similar words.

Accordingly, the goal of this experiment was to estimate whether some of the images misclassified by a model relying on TAGS features only could be corrected introducing

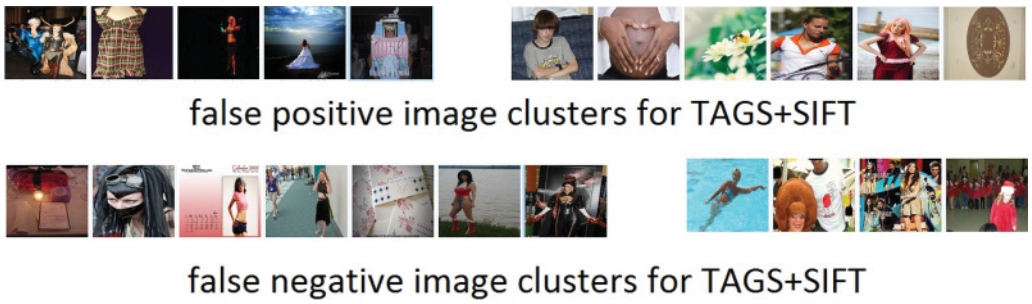

Fig. 12. False-positive and false-negative images of TAGS and SIFT+FACIAL.

SIFT and whether these errors are fewer in number but of the same images or whether the errors with SIFT+TAGS are due to a different set of images. Figure 12 reports the results of false positives and false negatives after running the image similarity analysis for the ensemble classifier of SIFT and TAGS. Since the prediction accuracy of SIFT and TAGS is better than TAGS, it was expected that the analysis would show a reduced number of images. However, the results show an interesting pattern. The image patterns that were observed by using TAGS as a classifier were not observed by using the ensembles classifier. The clustering of images revealed that the main error of classifying images using the ensembles classifier were because of unclear faces or background. In some cases, we observed images that could have been labeled otherwise based on human inspection. As such, the error patterns of the single classifier TAGS were mostly eliminated by the ensemble classifier, and the few remaining misclassified images are primarily due to some atypical content in the images itself (e.g., images with no objective sensitive content still labeled as private). This means that not only does an ensemble classifier performs accurate classification, but also that it is generally better at understanding visual patterns. For instance, images representing women or artwork were not misclassified in large numbers. The clusters in Figure 12 show a completely different error pattern compared to TAGS. As we return in Section 7, this analysis seems to warrant the need of more investigation. For instance, more advanced ensemble models that include complex visual features could further help in improving overall accuracy without having to include metadata.

5.3.1. Secondary Dataset Experiments. In order to verify whether our models are bound to a specific dataset, we sampled a new set of images from the Visual Sentiment Ontology [Borth et al. 2013] repository. The sampling was done by randomly selecting a URL from the complete list of about 45,000 images. Using the sampled set, we tested our best classifier, an 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 1,100 to 2,500. Figure 13 shows the models' prediction accuracy, computed using TAGS and TAGS+SIFT, across both the PicAlert and Visual Sentiment Ontology datasets. We make the following observations:

—The obtained accuracy is comparable with the accuracy observed for the models against the PicAlert dataset.
—The overall pattern of increase and decrease in predication accuracy with a change in size of the training set was consistent across both datasets.
—The performance of individual features is not preserved. 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

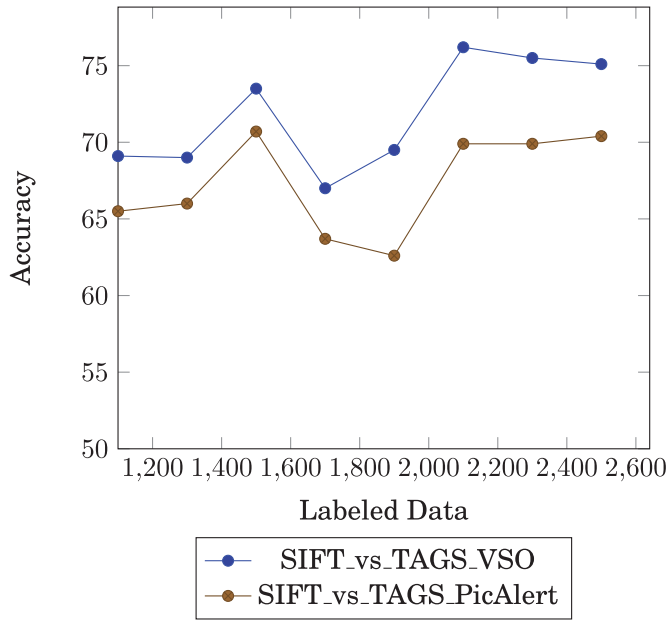

Fig. 13. Comparison between performance for models on the PicAlert and Visual Sentiment Ontology datasets.

availability of high-quality tags available in the first dataset. Nevertheless, we note that the final prediction accuracies across different sizes of datasets are within a range of 2–3% when we compare the results from the two 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, distinguish the possible access privileges. Accordingly, on 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).

6.1. Learning Models

Our privacy setting 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. 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 commentable and downloadable, it would be possible only if the image is viewable to the users in the same level of privacy.

Therefore, intuitively, if we were to consider a classification model that classifies the images according to three labels independently, we may obtain very disparate results for each class. To account for these types of inter-relations of categories for classifying unlabeled images, we apply the Chained Classifiers model [Read et al. 2011]. Chained classifiers were developed to capture the dependency between categories in a multi-category dataset. This 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 [Read et al. 2011] involves three transformations, one for each label. In a sense, the 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, $\{c_1, c_2, \ldots, c_N\}$, 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 to increase the predictive accuracy. As an example of a correlation in our use case, the *comment* label has the predicted value of the *view* label in the attribute space. Intuitively, an image can only be commented on if it can be viewed.

6.2. Experimental Results

We again relied on the PicAlert dataset for these experiments. We sampled 4,500 images and used the same features set as in our previous experiments. For labeling purposes, we used AMT. The quality of the 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. Our final dataset included 4,427 images.

Workers saw an image and were asked to select the most suitable privacy option they would pick based on their privacy preferences, assuming such an image were to be displayed on a social networking site. The question wording was as follows: "Assume you have taken this photo, and you are about to upload it on your favorite social networking or content sharing site (e.g., Facebook, Flickr, Google+, Picasa). Assume it describes something related to your life. Please tell us your privacy preferences, that is, by answering the following simple questions." Questions followed (see Section 6.1 for the options) per every available option (view, comment, download).

As a result of this labeling effort, the distribution of labels for every image is as follows: Forty-two percent of data is labeled as "everyone," and family and friends constitute about 33% of the data. Finally, the most private labels are assigned to 21% of the image dataset. Note that, as expected, almost 50% of the images are labeled as public. This is acceptable, as the images were actually taken from public accounts of the Flickr site.

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 that append 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. When referring to chain classifiers, in our experiments it is important to note the following. To ensure best performance, the order in which classifiers are chained with one another is important. That is, should the output of the "download" be the additional attribute in the next level of classification?

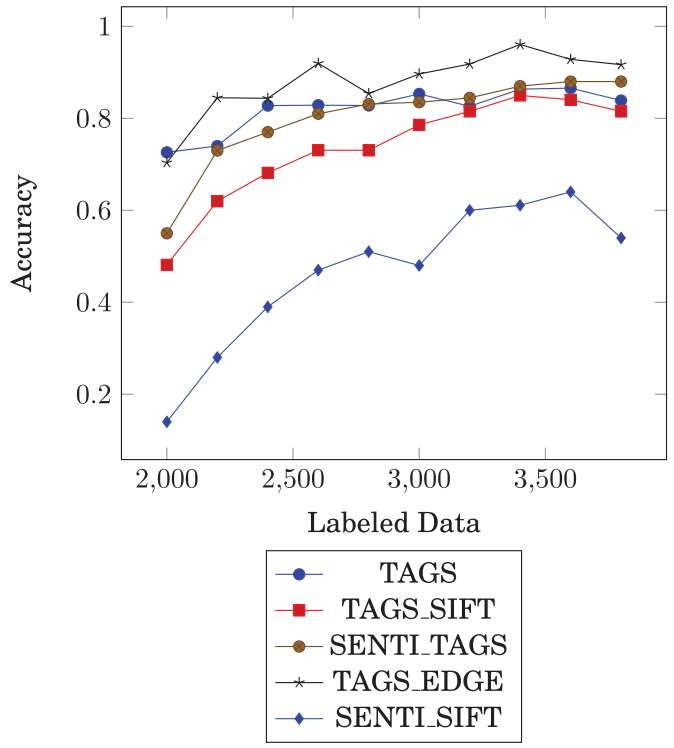

Fig. 14. Multi-option privacy settings Baseline.

To make that decision, we carried out the following simple evaluation. Intuitively, one may be able to comment on an image on viewing it, and not vice versa. Therefore, there may be a weak implication on "comment→view," in that if a group of users is allowed to comment on an image, at least the same users (if no more) are allowed to view it. Similar considerations apply for the implication "download→comment." Our experiments confirm this expected order. For instance, in an experiment with 4,000 images (3,500 training and 500 testing), using TAGS as the only feature, we obtained the following results. With the order of {comment, view, download}, our performance was 0.77 precision and 0.0754 recall. Conversely, with an order of {view, comment, download}, we obtain 0.62 precision and 0.606 recall, respectively. Baseline results showed 0.64 precision and 0.604 recall. Accordingly, the order of {view, comment, download } was maintained through the chain classifiers. Our results for both baseline and chained classifier models are reported in Figure 14 and in Figure 15. Note that in the figure, we reported results for the three features that performed best in our binary classification, namely SIFT, TAGS, and EDGES. We also include experimental results for sentiment features. We increased the size of the training data, ranging from 2,000 to 4,000, keeping 500 as the size of the test data.

A few interesting observations can be made. First, 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%.

We also note that the model achieves an accuracy reaching around 88% in the absence of TAGS. Specifically, the combination of sentiment features plus edge yields an accuracy of 88.94% (non-statistically significant). Because of the small difference in

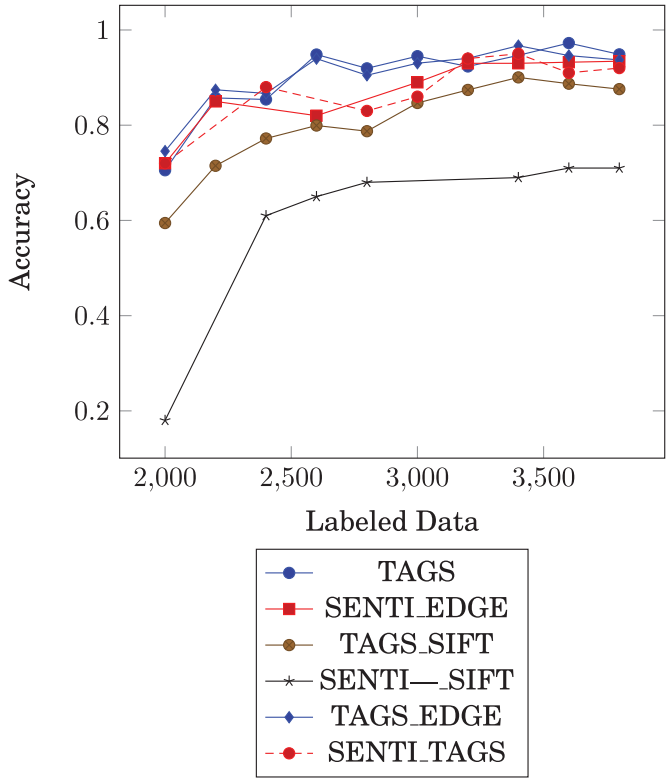

Fig. 15. Multi-option privacy settings.

performance, it is clear that the introduction of sentiment feature, while interesting, is not as informative as expected, certainly with respect to TAGS features.

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. For instance, in the baseline model, using TAGS alone reaches maximum accuracy of 86.5% accuracy when 3,600 images are used for training. With chained classifiers, the same ratio of training and testing data brings the performance up to 96.2%. Similar considerations may be made for other classes. This result confirms that chained classifiers are useful in capturing the correlation between the class labels and result in higher prediction accuracy.

7. DISCUSSION AND CONCLUSION

In this article, we presented an extensive study investigating models for inferring the degree of privacy of user-uploaded online images. We studied how features extracted from an image visual content and from user-applied metadata can inform possible privacy settings for the same image. We considered both the case of binary privacy settings (i.e. public/private), as well as the case of more complex, multi-class, privacy options. Our analysis provides us with several insights on images content, their relation with privacy, and on the ability to create a model that accurately describes what the users' privacy patterns actually are.

First, we note that, overall, we confirm our initial hypothesis: It is possible to capture with a certain degree of accuracy the "private" nature of an image, based solely on its

visual content. This is true not only for images that are not managed by a single individual, but also for images where privacy is defined as a "collective" notion and where the labels (i.e. private/public) reflect the overall preferences of a large number of users. As expected, most private images are mostly linked to suggestive content, the presence of children, and other life events that one may perceive as generally private. To capture these trends, our analysis shows that, with respect to visual features, properly trained SIFT features seem to carry strong predictive power. Colors, faces, and shapes or edge recognition are not sufficient to capture the complex nuances of a private image, although they work well together—even simply concatenated with one another—and achieve stronger predictive power than in isolation. In this respect, we also tested the hypothesis that negative sentiments may be linked with private images (and vice versa), in that online sites tend to primarily portray happy images when public, leaving negative images for a smaller audience who may better sympathize with those feelings. We note, however, that we could only partially validate this hypothesis, and more investigation is needed. Adding the complex "sentiment" feature seems to help only in specific model combinations. In general, the addition of keywords is substantially more important in improving the performance of a prediction, regardless of the specific model combination being considered.

Our work is yet just a tipping point in addressing the complex issue of online photo privacy. As images increasingly become a main form of communication and self-disclosure, more efficient models are to be investigated. Moving forward, to provide even more accurate models than the ones proposed in this work, we plan to extend our work along the following dimensions.

First, given the strong performance of TAGS, we would like to incorporate additional textual metadata into the models to further the performance of this class of features.

Second, as anticipated in Section 5.2.5, we would like to further explore more sophisticated visual models and their roles in the context of complex privacy settings. In particular, we are currently exploring using neural networks to better define the weights and relative importance of the features used for our analysis, possibly to improve the overall performance of our model in absence of tags. Early experiments show a strong predictive power of models built on convolutionary neural networks, although the complexity linked with training neural network models makes it challenging to generalize any result.

Finally, we plan to extend our dataset with possibly user-owned images for personalized models. This will allow us to verify whether the "collective" notion of private images extracted by these models could suit personalized privacy choices.

REFERENCES

Shane Ahern, Dean Eckles, Nathaniel S. Good, Simon King, Mor Naaman, and Rahul Nair. 2007. Over-exposed? Privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*. ACM, New York, NY, 357–366. DOI:http://dx.doi.org/10.1145/1240624.1240683

Eytan Adar Alessandra Mazzia, Kristen LeFevre. 2011. UM Tech Report #CSE-TR-570-11.

Morgan Ames and Mor Naaman. 2007. 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)*. 971–980. DOI:http://dx.doi.org/10.1145/1240624.1240772

Andrew Besmer and Heather Lipford. 2009. Tagged photos: Concerns, perceptions, and protections. In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY, 4585–4590. DOI:http://dx.doi.org/10.1145/1520340.1520704

Social Discovery Blog. 2012. Pin or not to Pin: An Inside Look. Retrieved from http://blog.socialdiscovery.org/tag/statistics/.

Joseph Bonneau, Jonathan Anderson, and Luke Church. 2009a. Privacy suites: Shared privacy for social networks. In *Proceedings of the Symposium on Usable Privacy and Security*.

Joseph Bonneau, Jonathan Anderson, and George Danezis. 2009b. Prying data out of a social network. In *ASONAM: Proceedings of the International Conference on Advances in Social Network Analysis and Mining*. 249–254.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 223–232.

Bullguard. 2014. Privacy violations, the dark side of social media. Retrieved from http://www.bullguard.com/bullguard-security-center/internet-security/social-media-dangers/privacy-violations-in-social-media.aspx.

O. Chapelle, P. Haffner, and V. N. Vapnik. 1999. Support vector machines for histogram-based image classification. *IEEE. Trans. Neur. Netw.* 10, 5 (1999), 1055–1064.

S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux. 2009. Img(Rummager): An interactive content based image retrieval system. In *Proceedings of the 2nd International Workshop on Similarity Search and Applications (SISAP'09)*. 151–153.

Gorrell P. Cheek and Mohamed Shehab. 2012. Policy-by-example for online social networks. In *17th ACM Symposium on Access Control Models and Technologies (SACMAT'12)*. ACM, New York, NY, 23–32.

Hong-Ming Chen, Ming-Hsiu Chang, Ping-Chieh Chang, Ming-Chun Tien, Winston H. Hsu, and Ja-Ling Wu. 2008. SheepDog: Group and tag recommendation for flickr photos by automatic search-based learning. In *Proceeding of the 16th ACM International Conference on Multimedia (MM'08)*. ACM, New York, NY, 737–740. DOI:http://dx.doi.org/10.1145/1459359.1459473

Munmun De Choudhury, Hari Sundaram, Yu-Ru Lin, Ajita John, and Dorée Duncan Seligmann. 2009. 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'09)*. IEEE, 1238–1241.

R. da Silva Torres and A. X. Falcão. 2006. Content-based image retrieval: Theory and applications. *Rev. Inf. Teór. Apl.* 2, 13 (2006), 161–185.

R. Datta, D. Joshi, J. Li, and J. Z. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40, 2 (2008), 5.

Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. 2010. 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)*. Springer-Verlag, Berlin, 71–84. Retrieved from http://portal.acm.org/citation.cfm?id=1888150.1888157

Lujun Fang and Kristen LeFevre. 2010. Privacy wizards for social networking sites. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY, 351–360.

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

X. He, W. Y. Ma, O. King, M. Li, and H. Zhang. 2002. Learning and inferring a semantic space from user's relevance feedback for image retrieval. In *Proceedings of the 10th ACM International Conference on Multimedia*. ACM, 343–346.

Benjamin Henne, Christian Szongott, and Matthew Smith. 2013. SnapMe if you can: Privacy threats of other peoples' geo-tagged media and what we can do about it. In *Proceedings of the 6th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. ACM, 95–106.

Kelly Jackson Higgins. 2010. Social Networks For Patients Stir Privacy, Security Worries. Retrieved from http://www.darkreading.com/authentication/167901072/security/privacy/227500908/social-networks-for-patients-stir-privacy-security-worries.html.

Simon Jones and Eamonn O'Neill. 2011. Contextual dynamics of group-based sharing decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, 1777–1786. DOI:http://dx.doi.org/10.1145/1978942.1979200

Peter F. Klemperer, Yuan Liang, Michelle L. Mazurek, Manya Sleeper, Blase Ur, Lujo Bauer, Lorrie Faith Cranor, Nitin Gupta, and Michael K. Reiter. 2012. Tag, you can see it! Using tags for access control in photo sharing. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'12)*. ACM. Retrieved from http://www.ece.cmu.edu/~lbauer/papers/2012/chi2012-tags.pdf.

Jinaga Latha. 2011. Java Content Based Image Retrieval. Retrieved from https://code.google.com/p/jcbir/.

Kun Liu and Evimaria Terzi. 2010. A framework for computing the privacy scores of users in online social networks. *ACM Trans. Knowl. Discov. Data* 5, Article 6 (Dec. 2010), 30 pages. Issue 1.

D. G. Lowe. 2004a. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 2 (2004), 91–110.

David G. Lowe. 2004b. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 2 (Nov. 2004), 91–110. DOI:http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

Michelle Madejski, Maritza Johnson, and Steven M. Bellovin. 2012. A study of privacy settings errors in an online social network. In *Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE, 340–345.

Andrew D. Miller and W. Keith Edwards. 2007. Give and take: A study of consumer photo-sharing culture and practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*. ACM, New York, NY, 347–356. DOI: http://dx.doi.org/10.1145/1240624.1240682

Wing W. Y. Ng, Andres Dorado, Daniel S. Yeung, Witold Pedrycz, and Ebroul Izquierdo. 2007. Image classification with the use of radial basis function neural networks and the minimization of the localized generalization error. *Pattern Recogn.* 40, 1 (2007), 19–32. DOI: http://dx.doi.org/10.1016/j.patcog.2006.07.002

Anon Plangprasopchok and Kristina Lerman. 2007. Exploiting social annotation for automatic resource discovery. *CoRR* abs/0704.1675 (2007).

Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. 2011. Automatic creation of photo books from stories in social media. *ACM Trans. Multimedia Comput. Commun. Appl.* 7S, 1, Article 27 (Nov. 2011), 18 pages.

Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. 2012. 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)*. ACM, New York, NY, Article 11, 8 pages.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD Record*, Vol. 29. ACM, 427–438.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning* 85, 3 (2011), 333.

Jose San Pedro and Stefan Siersdorfer. 2009. Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, New York, NY, 771–780. DOI: http://dx.doi.org/10.1145/1526709.1526813

Neela Sawant. 2011. Modeling tagged photos for automatic image annotation. In *Proceedings of the 19th ACM International Conference on Multimedia (MM'11)*. ACM, New York, NY, 865–866.

Neela Sawant, Jia Li, and James Ze Wang. 2011. Automatic image semantic interpretation using social action and tagging data. *Multimedia Tools Appl.* 51, 1 (2011), 213–246.

Josef Sivic and Andrew Zisserman. 2003. Video google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*. IEEE, 1470–1477.

Anna Cinzia Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2014. Analyzing images? Privacy for the modern web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. ACM, 136–147.

Anna Cinzia Squicciarini, Smitha Sundareswaran, Dan Lin, and Joshua Wede. 2011. A3P: Adaptive policy prediction for shared images over popular content sharing sites. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*. ACM, 261–270.

Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Shaohui Liu. 2009. Photo assessment based on computational visual attention model. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*. ACM, New York, NY, 541–544. DOI: http://dx.doi.org/10.1145/1631272.1631351

H. Sundaram, L. Xie, M. De Choudhury, Y. R. Lin, and A. Natsev. 2012. Multimedia semantics: Interactions between content and community. *Proc. IEEE* 100, 9 (2012), 2737–2758.

Aditya Vailaya, Anil Jain, and Hong Jiang Zhang. 1998. On image classification: City images vs. landscapes. *Pattern Recogn.* 31, 12 (1998), 1921–1935.

Nitya Vyas, Anna Cinzia Squicciarini, Chih-Cheng Chang, and Danfeng Yao. 2009. Towards automatic privacy management in web 2.0 with semantic analysis on annotations. In *CollaborateCom*. 1–10.

Chong Wang, David M. Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, 1903–1910.

Haitao Xu, Haining Wang, and Angelos Stavrou. 2015. Privacy risk assessment on online photos. In *Research in Attacks, Intrusions, and Defenses*. Springer, 427–447.

Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In *Proc. of ACM Workshop on Multimedia Information Retrieval*. ACM, 197–206.

Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, and Ming Ouhyoung. 2010. Personalized photograph ranking and selection system. In *Proceedings of the International Conference on Multimedia (MM'10)*. ACM, New York, NY, 211–220. DOI: http://dx.doi.org/10.1145/1873951.1873963

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

Jie Yu, Xin Jin, Jiawei Han, and Jiebo Luo. 2010. Social group suggestion from user image collections. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY, 1215–1216. DOI:http://dx.doi.org/10.1145/1772690.1772881

J. Yu, D. Joshi, and J. Luo. 2009. Connecting people in photo-sharing sites by photo content and user annotations. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME'09)*. IEEE, 1464–1467.

Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. 2012. Privacy-aware image classification and search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. ACM, New York, NY, 35–44. http://doi.acm.org/10.1145/2348283.2348292

Nan Zheng, Qiudan Li, Shengcai Liao, and Leiming Zhang. 2010. Which photo groups should I choose? A comparative study of recommendation algorithms in flickr. *J. Inf. Sci.* 36 (Dec. 2010), 733–750. Issue 6.

Jinfeng Zhuang and Steven C. H. Hoi. 2010. Non-parametric kernel ranking approach for social image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'10)*. ACM, New York, NY, 26–33. DOI:http://dx.doi.org/10.1145/1816041.1816047

Received May 2015; revised September 2016; accepted October 2016