On the Use of "Deep" Features for Online Image Sharing

Ashwini Tonge and Cornelia Caragea
Computer Science, Kansas State University
Manhattan, Kansas
atonge@ksu.edu,ccaragea@ksu.edu

ABSTRACT

Online image sharing in social networking sites such as Facebook, Flickr, and Instagram can lead to unwanted disclosure and privacy violations, when privacy settings are used inappropriately. Despite that social networking sites allow users to set their privacy preferences, this can be cumbersome for the vast majority of users. In this paper, we explore privacy prediction models for social media that can automatically identify private (or sensitive) content from images, before they are shared online, in order to help protect users' privacy in social media. More precisely, we study "deep" visual features that are extracted from various layers of a pre-trained deep Convolutional Neural Network (CNN) as well as "deep" image tags generated from the CNN. Experimental results on a Flickr dataset of thousands of images show that the deep visual features and deep image tags can successfully identify images' private content and substantially outperform previous models for this task.

CCS CONCEPTS

• **Security and privacy → Social network security and privacy**;

KEYWORDS

image privacy; deep visual features; privacy setting prediction

ACM Reference Format:
Ashwini Tonge and Cornelia Caragea. 2018. On the Use of "Deep" Features for Online Image Sharing. In *The 2018 Web Conference Companion, April23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3191572

1 INTRODUCTION

The rapid increase in multi-media sharing through social networking sites such as Facebook, Flickr, and Instagram can cause potential threats to users' privacy, when privacy settings are used inappropriately [1]. Many users quickly share private images about themselves, their family and friends, but they rarely change the default privacy settings, which could jeopardize their privacy [22]. These shared images can potentially reveal a user's personal and social habits. Furthermore, the smartphones facilitate the exchange of information virtually at any time with people all around the world. A study by the Pew Reserch center [10] of the social networking sites users regret the posted content. Users' privacy is recognized as a concern by social networking sites researchers as well. For example, the Director of AI Research at Facebook, LeCun [8] urges the development of a digital assistant, to warn people about sensitive content while uploading embarrassing photos. Thus, in order to avoid privacy violations and protect users' shared content in social media, it has become critical to develop automated privacy-aware models that can accurately detect private (or sensitive) content from images before they are shared online.

(a) Private

(b) Public

Figure 1: Examples of private and public images.

A naive rule-based classifier that classifies an image as private if it contains people does not work well in a real-world scenario. For example, Laxton et al. [7] described a "tele-duplication attack" that allows an adversary to create a physical key duplicate simply from an image. The rule-based model will fail to predict the image of a key as consisting of private (or sensitive) content, which needs to be protected. Figure 1 shows examples of images having *private* or *public* content, from a publicly available dataset [22].

Prior works explored binary prediction models of image privacy based on user tags and image content features such as SIFT (Scale Invariant Feature Transform) and RGB (Red Green Blue) [17, 22]. Authors found that SIFT features and user tags are informative for the task of classifying images as *private* or *public*. Yet, as images' tags are at the sole discretion of users, they tend to be noisy and incomplete [18]. Recently, due to the success of object recognition from images using Convolutional Neural Networks (CNNs) [6], researchers started to investigate privacy frameworks based on CNNs [19]. However, automatically identifying private content is inherently difficult because it requires an in-depth "understanding" of the visual content of the image. Additionally, the task is very subjective, depending on factors such as users' personalities and their privacy awareness. Recently, Zhong et al. [23] discussed challenges faced by both generic and personalized models for image privacy classification. Specifically, they highlight that generic privacy patterns do not capture well an individual's sharing behavior, whereas personalized models generally require large amounts of user data to learn reliable models, and are time and space consuming to train

WWW '18 Companion, April23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191572

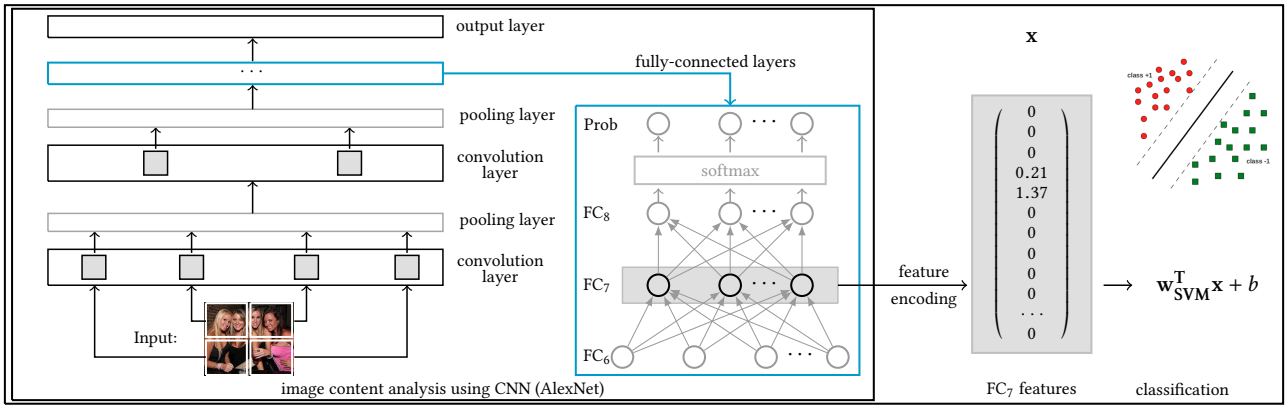

Figure 2: Deep features: CNN is used to extract deep visual features and deep image tags for input images.

and store models for each user. We recognize that progress should be made on both directions to improve hybrid approaches of generic and personalized models. Thus, in this paper, we aim at identifying a set of generic privacy patterns, i.e., "deep" features that have the highest discriminative power for privacy prediction.

Contributions. We present an analysis of various "deep" feature representations for image privacy prediction (i.e., for predicting the class of an image as *private* or *public*). Unlike previous works, we explore features that can be directly obtained from a pre-trained object CNN for privacy prediction. Specifically, we use deep feature representations corresponding to the output of fully-connected layers of a CNN, pre-trained on ImageNet [12], as well as the probability distribution over the object categories obtained from the last layer of the network via softmax. Since the set of user tags may be incomplete and noisy, unlike previous works, we leverage CNNs for automatically generating "deep tags" that correspond to the top-ranked probabilities obtained from the probability distribution over the 1,000 object categories. These tags can also provide the relevant cues for privacy-aware image retrieval [22].

We evaluate the performance of the "deep" features (extracted from AlexNet [6]) on a subset of the PicAlert dataset of Flickr images given by Zerr et al. [22], labeled as private or public. We empirically show that learning models trained on deep features for privacy prediction outperform strong baselines such as those trained on hierarchical deep features, SIFT, GIST (global image descriptors) and user tags. We also show that deep features provide improved performance for the private class as compared to baseline approaches. Moreover, the results show that the deep tags yield better performing models as compared to user tags and the combination of deep and user tags outperforms both set of tags.

2 RELATED WORK

Buschek et al. [2] presented an approach to assigning privacy to shared images using metadata (location, time, shot details) and visual features (faces, colors, edges). Zerr et al. [22] proposed privacy-aware image classification, and learned classifiers on Flickr photos. Authors considered user-annotated tags and visual features such as color histograms, faces, edge-direction coherence, and SIFT for the privacy classification task and found that SIFT has a high discriminative power for image privacy detection. Consistent with Zerr et al. [22], Squicciarini et al. [17] also found that SIFT and user-annotated tags work best for predicting privacy of users' images. Given the recent success of CNNs for various image-related tasks [3, 5, 6, 13, 14], Tran et al. [19] investigated CNNs for privacy prediction and showed improved performance compared with visual features such as SIFT [9] and GIST [11] (this approach is one of our strong baselines). Spyromitros-Xioufis et al. [15] explored features extracted from CNNs to provide more accurate personalized privacy classification. Yu et al. [21] adopted CNNs to achieve semantic image segmentation and also learned object-privacy relatedness to identify sensitive objects.

3 IMAGE PRIVACY PREDICTION

The privacy of an image can be determined by the presence of one or more objects described by the visual content and the description associated with it in the form of tags.

Problem Statement: Given an image to be uploaded online, the task is to classify it into one of the two classes: *private* or *public*, i.e., consisting of private or public content, respectively.

Next, we describe the features used in the classification.

Feature Extraction: We extract "deep" features from images using AlexNet CNN [6] pre-trained on the ILSVRC-2012 object classification subset of the ImageNet dataset [12]. AlexNet implements an eight-layer feed-forward neural network in which first five layers consist of interleaved convolution and pooling layers, and top three layers consist of fully-connected (FC) layers. The convolution layers represent high-level features of images, whereas the FC layers give the non-linear combination of the features in the layers below. A probability (prob) layer obtained by applying the softmax function to the input from the previous FC layer, and finally the output layer, which outputs the probabilities of the objects in the input image. This is illustrated in Figure 2. The reason for using features derived from a pre-trained network is that the sensitive content is limited for model training and training or fine-tuning a deep network requires a large amount of privacy data.

Deep Visual Features: We extracted deep visual features from the FC layers, which are referred as FC_6, FC_7, and FC_8, and from

the "prob" layer (the cyan block in Figure 2). The dimensions of FC_6, FC_7, and FC_8 are 4096, 4096 and 1000, respectively. The "prob" layer produces a probability distribution over $c = 1000$ object categories for the input image using softmax function and can be defined as: $P(y = c|\mathbf{z}) = \frac{exp(z_k)}{\sum_j exp(z_j)}$ where, $\mathbf{z}$ is the output of the FC_8 layer.

Deep Image Tags: It is interesting to mention that not all images on social networking sites have tags or the set of tags is very sparse [18]. Thus, we use an automatic annotation technique to derive tags for images based on their visual content. For automatic image annotation, we predict the top K object categories for an input image $\mathbf{x}$ from the probability distribution extracted from the CNN. we obtain deep tags such as "Maillot," "Wig," "Brassiere," "Bra," "Miniskirt" for the picture in Figure 2 (note that only top $K = 5$ deep tags are obtained). However, important tags such as "people" and "women" are not included. This is because the 1,000 object categories used for training do not contain these tags.

4 DATASET AND EVALUATION SETTINGS

We trained and evaluated models based on deep features on a subset of 32,000 Flickr images sampled from the PicAlert dataset, made available by Zerr et al. [22]. PicAlert consists of Flickr images on various subjects, which are manually labeled as *private* or *public* by external viewers. In our experiments, the 32,000 images are split into **Train** and **Test** sets of 10,000 and 22,000 images, respectively. We consider a higher number of test images (compared to Training images) to evaluate the "deep" features on a large set of unseen images for limited number of training images. Each experiment was repeated five times with a different train/test split and the micro averaged results are presented across these five runs. The public and private images are in the ratio of 3:1 in both train and test.

Evaluation Setting. To evaluate the deep features, we used the Support Vector Machine (SVM) classifier implemented in Weka and chose the hyper-parameters that gave the best performance on the **Train** set using 10-fold cross-validation (CV). We experimented with $C = \{0.001, 0.01, 1.0, \cdots, 10.0\}$, kernels: Polynomial and RBF, the γ parameter in RBF, and the degree d of a polynomial. Hyper-parameters shown in all result tables follow the format: "R/P,C,γ/d" where "R" denotes "RBF" and "P" denotes "Polynomial."

5 EXPERIMENTS AND RESULTS

We present the experimental evaluation of the deep features. We compare the performance of the models trained on deep visual features with the models trained on baseline visual features for privacy prediction. Earlier user tags performed well for privacy prediction [17, 22], and hence, we examine the quality of tag features using both user annotated tags and automatically annotated (deep) tags.

5.1 Results for Deep Visual Features

Experimental design: We wish to identify the most promising visual features from the set of deep features that have the highest discriminative ability for privacy classes. To achieve this, we first compare the deep visual features among each other. We then compare the performance of models based on deep visual features with several baselines that we described below.

Baselines. Tran et al. [19] proposed PCNH, a privacy CNN-based framework, that combines features obtained from two architectures:

Features	H-Param	Acc %	F1	Prec	Re
#1 Deep visual features					
FC_6	R,1.0,0.05	85.49	0.844	0.847	0.855
FC_7	R,2.0,0.01	**85.83**	**0.851**	**0.851**	**0.858**
FC_8	R,1.0,0.05	85.80	**0.851**	**0.851**	**0.858**
Prob	R,5.0,1.0	83.18	0.824	0.822	0.832
#2 Hierarchical Deep Features [19]					
PCNH	–	84.21	0.833	0.832	0.842
#3 SIFT/GIST [16, 17, 22]					
SIFT	P,1.0,2.0	77.31	0.674	0.598	0.773
GIST	R,0.001,0.5	77.33	0.674	0.598	0.773
SIFT+GIST	R,0.05,0.5	72.67	0.704	0.691	0.727
#4 Rule-based models					
Rule-1	–	77.35	0.683	0.694	0.672
Rule-2	–	77.93	0.673	0.704	0.644

Table 1: Deep visual features vs. Baselines

one that extracts convolutional features, and another that extracts object features. The Object CNN is a deep network of 11 layers obtained by appending three FC layers of size 512, 512, 24 at the end of the FC layer of AlexNet. The PCNH framework is first trained on the ImageNet dataset and then fine-tuned on a small privacy dataset. As images' privacy greatly depends on the objects in images, we believe that the features controlling the distinct attributes of the objects obtained through the higher number of neurons (4096 neurons in FC_7 of AlexNet) can better approximate the privacy function compared with adding more non-linear layers (as in PCNH). The increase in the number of complex non-linear layers introduces more parameters to learn, and at the same time, with comparatively small amount of training data (PicAlert vs. ImageNet), can result in over-fitting. Moreover, training such a deep network on ImageNet and then fine-tuning on the privacy data significantly increases the processing power and time complexity. Furthermore, if new objects are added to the object dataset, the networks need to be retrained from scratch. Conversely, features derived from state-of-the-art CNN architectures can reduce the overhead of re-training and still achieve good performance for privacy prediction. Hence, we compare models trained on the "deep" features with the PCNH privacy framework, and consider the latter as our first baseline. Unlike Tran et al. [19] who used 800 images in their evaluation, we evaluate our models on a large set of images (22000) to validate the performance of the deep features for a large variety of image subjects. We regard classifiers trained on the best performing features between SIFT, GIST, and their combination as the second strong baseline. We also compare the performance of the deep features with two naive rule-based classifiers, which predict an image as *private* if it contains persons. Otherwise, the image is classified as *public*. For the first rule-based classifier, we detect front and profile faces by using Viola-Jones algorithm [20]. For the second rule-based classifier, we consider user tags such as "women," "men," "people."

For the deep visual features, we use the AlexNet pre-trained CNN implemented in CAFFE [4], which is an open-source framework for deep neural networks. We resize images in both **Train** and **Test** to the CAFFE convolutional neural net compatible size of 227×227 and encode each image using the three deep feature representations corresponding to the output of the layers FC_6, FC_7, FC_8, and "Prob," which is the probability distribution obtained from FC_8 via softmax.

Results: Table 1 shows results of the comparison (Precision, Recall, F1- Measure and Accuracy) of SVMs using each deep feature type extracted from AlexNet, FC_6, FC_7, FC_8, and "Prob," and the results of their comparison with the performance of baselines (i.e., SVMs trained using the baseline features), on **Test**. We can see from the table that the SVMs trained on FC_7 and FC_8 perform similarly, and the performance improves as we go from FC_6 to FC_7. This is because higher layers of the network capture high level feature descriptions of objects present in the image. We notice that all FC_6, FC_7, FC_8 deep features are able to achieve performance higher than 85% in terms of all compared measures. Note that a naive baseline which classifies every image as "public" obtains an accuracy of 75%. It is worth mentioning that "prob" features perform worse than the features extracted from the fully-connected layers. One possible explanation could be that squashing the values at the previous layer (FC_8 in AlexNet) through the softmax function, which yields the "prob" layer, produces a non-linearity that is less useful for SVM compared to the un-transformed values. The results of FC layers over the "prob" layer are statistically significant for p-values < 0.05.

Table 1 shows also that deep visual features FC_6, FC_7, FC_8 provide better feature representations than baseline visual features for privacy prediction. Precisely, the models obtained using deep visual features extracted from AlexNet outperform models trained on baseline features, PCNH, SIFT, GIST and SIFT + GIST. For example, F1-measure improves from 0.833 obtained by PCNH features to 0.851 obtained by FC_8. We achieve improvement in F1-measure as high as 15% over SIFT + GIST models, i.e., our second baselines. "Prob" features also perform better than SIFT + GIST. With a paired T-test, our improvements over the baseline approaches for F1-measure are statistically significant for p-values < 0.05. It is also interesting to note that rules based on facial features exhibit better performance than SIFT and GIST and suggest that features representing persons are helpful to predict private content of images. However, "deep" features outperform the rule-based models based on facial features by more than 10% in terms of all measures (see Table 1, #4 Rule-based models). Simple rule-based models will not suffice for this task and advanced AI technology such as deep learning is required.

Features	F1	Prec	Re
#1 Deep visual features			
FC_7	**0.642**	**0.752**	**0.56**
#2 Hierarchical Deep Features [19]			
PCNH	0.598	0.708	0.518
#3 SIFT/GIST [16, 17, 22]			
SIFT+GIST	0.27	0.343	0.223
#4 Rule-based models			
Rule-1	0.509	0.47	0.556
Rule-2	0.458	0.373	0.593

Table 2: Performance for "Private" class.

We also show the privacy prediction performance for "private" class in Table 2 to identify which features characterize the private class effectively as sharing private images on the Web with everyone is not desirable. We found that the SVM trained on AlexNet-based deep visual features obtain improved performance for the private class as compared with the SVM trained on the baseline features. Precisely, using the best-performing deep visual features FC_7, F1-measure for the private class improves from 0.598 obtained by PCNH to 0.642 obtained by FC_7.

Features	H-Param	Acc %	F1	Prec	Re
User Tags	R,2.0,0.05	81.73	0.789	0.803	0.817
DT	R,1.0,0.1	83.18	0.819	0.819	0.832
DT+UT	R,1.0,0.05	**84.59**	**0.833**	**0.837**	**0.846**

Table 3: Privacy prediction performance using tag features.

Next, we examine the quality of tag features and contrast the deep image tags with the user annotated tags.

5.2 Results for Deep Image Tags

Experimental design: We investigate the performance of SVMs on user tags and deep image tags for privacy prediction. We also examine the combination of user tags and deep tags, which captures different aspects of an image. Examples of user tags for the image in Figure 2 are: "Birthday Party," "Night Life," "People," etc. For the deep tags, we consider $K = 10$ as other K values did not yield higher results (in 10-fold CV over the Train set).

Results: Table 3 shows the results obtained from the experiments for tag features on the **Test** and compares the performance obtained using models trained on deep tags, user tags and their combination. In the table, "UT" represents user tags and "DT" represents deep tags. From the table, we notice that deep tags perform better than user tags, however, the combination of the two outperforms each one individually, the user tags and the deep tags. This can be justified by the fact that the user tags have some general tags, whereas deep tags contain some specific tags, which capture various aspects of the data. To see this, using only general tags can cause overlap in the two different privacy classes. For example, if we consider more general tags such as "clothes" instead of "swimsuit," then the tag can appear in both classes and hence will fail to differentiate between them. Similarly, if we would consider only very specific tags, the models may overfit and will not generalize well on unseen data.

6 CONCLUSION AND FUTURE WORK

In this paper, we explored AI technology, i.e., deep features extracted from various CNN layers, for image privacy classification. Our results show that the deep visual features corresponding to the fully-connected layers of the AlexNet CNN outperform those corresponding to the "prob" layer. We also examined user annotated tags and deep tags (generated from the "prob" layer) and found that the combination of both the tags outperforms individual sets of tags. In addition, models trained on deep features yield improvement in performance over several baselines. The result of our classification task is expected to aid other very practical applications. For example, a law enforcement agent who needs to review digital evidence on a suspected equipment to detect sensitive content in images and videos, e.g., child pornography. The learning models developed here can be used to filter or narrow down the number of images and videos having sensitive or private content before other more sophisticated approaches can be applied to the data. In future, other CNN architectures can be explored for privacy prediction. Also, user tags can be extracted from description, and comment to obtain additional information about the image.

7 ACKNOWLEDGMENTS

This research is supported in part by the NSF award #1421970.

REFERENCES

[1] 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 (CHI '07)*. ACM, New York, NY, USA, 357–366.

[2] Daniel Buschek, Moritz Bader, Emanuel von Zezschwitz, and Alexander De Luca. 2015. Automatic Privacy Classification of Personal Photos. In *Human-Computer Interaction - INTERACT 2015*. Vol. 9297. 428–435.

[3] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. 2013. Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (August 2013).

[4] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the ACM International Conference on Multimedia*. 675–678.

[5] Sergey Karayev, Aaron Hertzmann, Holger Winnemoeller, Aseem Agarwala, and Trevor Darrell. 2013. Recognizing Image Style. *CoRR* abs/1311.3715 (2013).

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105.

[7] Benjamin Laxton, Kai Wang, and Stefan Savage. 2008. Reconsidering Physical Key Secrecy: Teleduplication via Optical Decoding. In *Proceedings of the 15th ACM Conference on Computer and Communications Security (CCS '08)*. ACM, 469–478.

[8] Yann LeCun. 2017. Facebook Envisions AI That Keeps You From Uploading Embarrassing Pics. https://www.wired.com/2014/12/fb/all/1. (2017). [Online; accessed 12-April-2017].

[9] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV* 60, 2 (Nov. 2004), 91–110.

[10] MARY MADDEN. 2012. Privacy management on social media sites. http://www.pewinternet.org/2012/02/24/privacy-management-on-social-media-sites. (2012). [Online; accessed 12-November-2017].

[11] Aude Oliva and Antonio Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV* 42, 3 (May 2001), 145–175.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* (April 2015), 1–42.

[13] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. 2014. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *ICLR 2014*. CBLS.

[14] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann Lecun. 2013. Pedestrian Detection with Unsupervised Multi-stage Feature Learning. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. 3626–3633.

[15] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Adrian Popescu, and Yiannis Kompatsiaris. 2016. Personalized Privacy-aware Image Classification. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16)*. ACM, New York, NY, USA, 71–78.

[16] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2017. Toward Automated Online Photo Privacy. *ACM Trans. Web* 11, 1, Article 2 (April 2017).

[17] Anna C. 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 (HT '14)*. ACM, New York, NY, USA, 136–147.

[18] Hari Sundaram, Lexing Xie, Munmun De Choudhury, Yu-Ru Lin, and Apostol Natsev. 2012. Multimedia Semantics: Interactions Between Content and Community. *IEEE* (2012).

[19] Lam Tran, Deguang Kong, Hongxia Jin, and Ji Liu. 2016. Privacy-CNH: A Framework to Detect Photo Privacy with Convolutional Neural Network Using Hierarchical Features. In *Proceedings of the Thirtieth AAAI Conference*. 1317–1323.

[20] Paul Viola and Michael Jones. 2001. Robust Real-time Object Detection. In *International Journal of Computer Vision*.

[21] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. 2017. iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. *IEEE Trans. Inf. Forensic Secur* 12, 5 (2017).

[22] 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*. ACM, NY, USA.

[23] Haoti Zhong, Anna Squicciarini, David Miller, and Cornelia Caragea. 2017. A Group-Based Personalized Model for Image Privacy Classification and Labeling. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3952–3958.