Pixel Privacy 2019:
Protecting Sensitive Scene Information in Images

Zhuoran Liu, Zhengyu Zhao, Martha Larson
Radboud University, Netherlands
{z.liu,z.zhao,m.larson}@cs.ru.nl

ABSTRACT

Pixel Privacy task focuses on the protection of user-uploaded multimedia data online. Specifically, it benchmarks image transformation algorithms that protect privacy-sensitive images against automatic inference. The image transformations should block automatic classifiers that infer sensitive scene categories and increase (or maintain) image visual appeal at the same time. The task in 2019 is to develop image transformations under the condition that all information of the attack model is available for transformation development. Under this *white-box* setting, the decreased accuracy of the attack model and the visual appeal of the protected images are considered for protection evaluation.

1 INTRODUCTION

The MediaEval Pixel Privacy task aims to promote the development of algorithms that protect the privacy-sensitive information of user-generated multimedia data online. To achieve this goal, participants are encouraged to develop *image transformation* algorithms that increase (or maintain) the visual appeal of images, while at the same time protecting privacy-sensitive information in the images. Ideally, users should find that the transformed images are interchangeable with the original image, for whatever purpose the original image was intended. The transformed images should be able to mislead automatic scene classifiers.

The task is motivated by the potential risk of the privacy-sensitive information implicit in user-generated data, which is accumulated by large social networks. Accumulated social media data can be misappropriated for commercial purposes that are not transparent to users [9]. Although algorithms [11, 13] have been developed to improve the the situation of privacy protection in multimedia online, users themselves still do not have many choices to control the information implicit in their own multimedia data. In addition, given the large amount of accumulated data, potential privacy risks could be aggravated by massive data breaches [9]. Privacy-sensitive information can be processed by automatic algorithms, allowing malicious actors to select potential victims as the target for specific crimes, a practice known as *cybercasing* [4]. For example, based on user-uploaded images, the trajectory of an individual can be calculated by geo-location prediction algorithms based on computer vision algorithms. This information can be exploited by a criminal to plan a burglary by only accessing the visual contents of these social photos. Combining mined information from different sources is also likely to aggravate online crimes, e.g., telecommunication fraud [1] or blackmail.

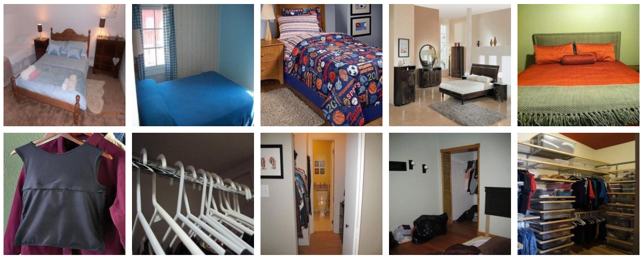

Figure 1: Examples of validation images in MediaEval Pixel Privacy task 2019. Images are randomly selected from category `bedroom` (top row) and `closet` (bottom row).

The Pixel Privacy task was introduced as a brave new task in the MediaEval Multimedia Evaluation Benchmark in 2018 [8]. The task focused on sensitive scene categories of social images, and required participants to protect images against an automatic scene classifier. Examples from the 2019 validation set are shown in Figure 1. In 2019, we again focus on the protection of sensitive scene categories and use the same basic task formulation and source data. The task has been refined in several ways in order to allow us to gain more insight from the results. In 2019, we retain the white-box setting, meaning that the attacking classifier is known and all information of the attack model is available for protection development. Also, we retain the untargeted setting, meaning that there is no particular target class into which the image must be misclassified. Instead, any misclassification counts as protection. The important change for this year is that the test set only contains images that the attacking classifier classifies correctly. For the purposes of evaluation, we find the images that the attacking classifier misclassifies to be less interesting because they can be considered to already be protected. Also, this year, we pay closer attention to the pipeline. Specifically, images are downsized before being fed into the attack classifier. Participants are required to protect the images in this downsized format, to control for the impact of the downsizing on the protection.

To achieve the goal of privacy protection and visual appeal improvement, researchers participating in the task may consider related work on different multimedia technologies. Image enhancement and style transfer [6] techniques can be exploited to increase visual appeal and protect privacy. Early work showed the basic ability of standard Instagram filters to block the inference of location information [3]. Last year, one participant paper [2] pursued a color harmony based enhancement approach, which focused on improving visual appeal and another investigated style transfer [10]. Image aesthetics assessment [14] and image quality assessment methods [5] could be helpful to control the visual quality of transformed images. Knowledge of adversarial examples in machine

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

learning can also be applied for privacy protection purpose [10, 11]. However, in 2018, participants did not fully exploit the white-box information, which we hope they will do in 2019.

2 TASK DEFINITION AND DATA

As stated above, the Pixel Privacy task 2019[1] focuses on the protection of privacy-sensitive scene category information of social images. A scene category can be understood to be the identity of the setting in which a photo was taken. Participants are asked to develop protection approaches on validation set to decrease the attack accuracy while increasing image appeal. Afterwards, these developed approaches can be applied on test set images, and the protected images are submitted for evaluation.

The task provides 60 privacy-sensitive categories chosen from the original 365 scene categories from Places365-Standard dataset [15], which were original introduced in 2018. The task data set is a subset of this dataset. The Places365-Standard dataset contains 1,803,460 training images and 365 scene categories. The number of images per category varies from 3,068 to 5,000. The attack algorithm is trained to detect all 365 categories. The attack classifier in the task is a PyTorch ResNet50[2] [7] classifier trained on the training set of the Places365-Standard dataset, as was also used in 2018.

A validation set (MEPP18val) is provided to allow participants to develop their image transformation algorithms. Figure 1 shows examples of image from the validation set. We also provide a test set (MEPP19test) to evaluate the performance of the transformation algorithms. MEPP18val contains 3000 images (50 from each of the 60 classes), while MEPP19test contains 600 images (around 10 from each of the 60 classes). Note that if the original images without modification can already block the attack model, no protection transformation is needed. Further, images which the original version is incorrectly classified by the classifier, may be correctly classified after transformation. To be able to measure protection performance without interference from these effects, MEPP19test is a subset of last year's test set (MEPP18test), and contains only the images that were correctly classified by the attack classifier.

Pixel Privacy task 2019 is a simplified version of social image privacy protection, and in particular, uses an untargeted white-box protection setting. Here, we provide some more details about what this means. The white-box setting is that all information of the attack model is available for image transformation development, which means the exact neural network architecture, pre-trained weights and related preprocessing details are available to participants. Untargeted setting defines no target categories for the protected images. In other words, if the predicted label is different from the ground truth then the protection is successful.

Preprocessing the transformed images may have strong influences on the protection performance evaluation. For this reason, in the task setting, no resizing and cropping are applied in the processing step. Normalization is the only preprocessing step carried out during evaluation. *Small images* (256*256) of Places365-Standard dataset are used as standard input, and they can be downloaded directly from the official website of Places data set[3].

For some creative image transformation ideas, it may not be feasible to develop fully automatic transformation algorithms. To leverage participants' creativity and explore unexpected new ways in improving the visual appeal, we also provide a special test set (MEPP19test_manual). It is a subset of test set and contains one image for each category. Manual image transformations can be applied on this special test set, and these images can also be submitted for evaluation.

3 EVALUATION

Participants submit the transformed test set for evaluation and each team can maximally submit five runs. Submitted images will be evaluated with respect to *protection* and *appeal*. The performance of transformation approaches with respect to protection is evaluated by measuring the drop of prediction accuracy of the *attack model*. Once the prediction accuracy has reached a certain level of protection, performance of transformations with respect to appeal will be carried out with an automatic aesthetics assessment algorithm. To this end, the automatic algorithm NIMA [14] trained on the AVA [12] dataset will be used for visual appeal evaluation, as was also done in 2018. This evaluation method aligns with practical needs from users for multimedia protective technologies.

In order to gain further insight in the appeal of images, we will perform further manual assessment on cases in which the NIMA scores for different protection algorithms diverge dramatically. We will select the images that have the highest variance of NIMA scores across runs submitted by all participating teams, and have these images inspected by a small panel of computer vision experts. The experts will choose the best and the worst examples from the pool of all protected versions. These examples will be qualitatively analyzed in order to gain further insight into the relative strengths and weaknesses of the different protection algorithms.

4 DISCUSSION AND OUTLOOK

One question remains is that whether changing the label of the image from the proper one to an arbitrary one is enough to help users hide their privacy-sensitive information? From Figure 1, we can imagine that if the label of an images is changed from bedroom to closet, the criminal may still be able to mine the information that this image is took from home. In this case, protection by changing the ground truth label to an arbitrary one is not enough. Another question is that in practical cases model information is not available, which means that the white-box setup may not be valid for image protection in real life.

Pixel Privacy task is a highly simplified task that defines how to protect users' multimedia data online in a user-controlled manner. In practice, the social multimedia data may have different types, e.g., text, video and speech data, and the threat models can be complicated too. The goal of the task is to provide a foundation upon which solutions addressing progressively more realistic versions of the problem may be developed in the future.

ACKNOWLEDGMENTS

This work is part of the Open Mind research program, financed by the Netherlands Organization for Scientific Research (NWO).

[1] https://github.com/multimediaeval/2019-Pixel-Privacy-Task

[2] http://places2.csail.mit.edu/models_places365/resnet50_places365.pth.tar

[3] http://places2.csail.mit.edu/download.html

REFERENCES

[1] 2017. Scammers still up to their tricks despite local efforts to stop them, China Daily, 21 July. (2017). http://www.chinadaily.com.cn/opinion/2017-07/21/content_30195232.htm, Online; accessed 8-Aug-2019.

[2] Simon Brugman, Maciej Wysokinski, and Martha Larson. 2018. MediaEval 2018 Pixel Privacy Task: Views on image enhancement. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*

[3] Jaeyoung Choi, Martha Larson, Xinchao Li, Kevin Li, Gerald Friedland, and Alan Hanjalic. 2017. The Geo-Privacy Bonus of Popular Photo Enhancements. In *ACM International Conference on Multimedia Retrieval (ICMR).* ACM, 84–92.

[4] Gerald Friedland and Robin Sommer. 2010. Cybercasing the Joint: On the Privacy Implications of Geo-tagging. In *Proceedings of the 5th USENIX Conference on Hot Topics in Security (HotSec).*

[5] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. 2017. Deep generative adversarial compression artifact removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 4826–4835.

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2414–2423.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 770–778.

[8] Martha Larson, Zhuoran Liu, Simon Brugman, and Zhengyu Zhao. 2018. Pixel Privacy: Increasing Image Appeal while Blocking Automatic Inference of Sensitive Scene Information. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*

[9] Sam Levin. 2017. Facebook Told Advertisers It Can Identify Teens Feeling 'Insecure' and 'Worthless', The Guardian, 1 May. (2017). https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens, Online; accessed 12-Jul-2019.

[10] Zhuoran Liu and Zhengyu Zhao. 2018. First Steps in Pixel Privacy: Exploring Deep Learning-based Image Enhancement against Large-scale Image Inference. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*

[11] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. 2019. Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR).* ACM, 306–314.

[12] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR).* IEEE, 2408–2415.

[13] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Adrian Popescu, and Yiannis Kompatsiaris. 2016. Personalized Privacy-aware Image Classification. In *ACM International Conference on Multimedia Retrieval (ICMR).* ACM, 71–78.

[14] Hossein Talebi and Peyman Milanfar. 2018. Nima: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[15] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.