Pixel Privacy: Increasing Image Appeal while Blocking Automatic Inference of Sensitive Scene Information

ABSTRACT

We introduce a new privacy task focused on images that users share online. The task benchmarks image transformation algorithms that are capable of blocking the ability of automatic classifiers to infer sensitive information in images. At the same time, the image transformations should maintain the original value of the image to the user who is sharing it, either by leaving it not obviously changed, or by enhancing it to increase its visual appeal. This year, the focus is on a set of 60 scene categories, selected from the Places365-Standard dataset, that can be considered privacy-sensitive.

1 INTRODUCTION

The objective of the MediaEval Pixel Privacy task is to promote the innovation of protective technologies that make it safer to share social multimedia online. Participants are provided with a set of images and asked to develop protective transformations that prevent the automatic detection of privacy-sensitive information contained in these images. In 2018, we use a subset of the Places365-Standard dataset associated with privacy-sensitive scene categories. The transformation algorithms may either leave the images not obviously changed (i.e., maintain the value of the image to the user who is sharing it) or else enhance their original appeal (i.e., increase the value of the image to the user who is sharing it). We especially encourage the development of transformation algorithms that enhance images, since we feel that users will be motivated to use them: it is easier to get excited about dressing up images than about taking precautionary measures that guard against intangible risks.

The Pixel Privacy task can be seen as an evolution of the MediaEval Visual Privacy task [1, 11] and DroneProtect task [2], which focused on protecting information in surveillance video from people watching the video. There are four key differences:

- We are interested in *social images* that users share online.
- We focus on image transformations that protect privacy-sensitive information against *automatic inference.*
- Consistent with this focus, we do not necessarily expect, nor do we require, that sensitive information is hidden from people who look at the images. (Our primary focus is protection against computer vision.)
- Our goal is irreversible protection (although results may also be relevant for reversible protection).

The motivation for the Pixel Privacy task is the growing concern about the information implicit in the user data that is shared online, and in particular, in the data accumulated by large social networks. Trust in social network platforms is necessary, but not enough. Events of recent years have made us realize how easily social network data can be misappropriated (e.g., [12]) or put to a use that is acceptable from the perspective of the social network company, but not from the perspective of users (e.g., [13]). Additionally, online data can be mined in order to search for victims, i.e., a so-called cybercasing attack [9]. A cybercasing attack is carried out by a malicious party who automatically searches through a large amount of social media in order to identify victims. For example, a criminal looking to identify houses to rob can make use of computer vision technology to rank images according to the probability that the user who shared them is currently traveling, and thus not at home. It has become clear that, in addition to trusting social networks, we need local technologies that provide users with more control over the information that can be inferred about them on the basis of the multimedia data that they share.

Researchers interested in tackling the Pixel Privacy task should apply their creativity to imagine which sorts of image enhancements users will find appealing. However, they should also dig deeply into the related work. In area of multimedia privacy, relevant work has been carried out on social images (e.g., [5]) and security video (e.g., [3, 6]). Beyond work on privacy, work on adversarial machine learning (e.g., [14, 15]) is also relevant.

2 TASK DEFINITION AND DATA

The larger goal of the Pixel Privacy task is to address the challenge of online multimedia privacy by creating user-controlled technologies (i.e., transformations that can be locally applied before sharing). Users must find the technologies easy and even fun to use. At the same time, the technologies must lower the risk of users suffering privacy violations due to the inference ability of computer vision algorithms that have been trained on large quantities of data. Potential violations include a range of different threats: being singled out as a victim (e.g., rich, not at home) for a harmful attack (e.g., break-in, blackmail) and being assigned by a commercial algorithm to a category (e.g., frequents slums, attends church) for the purposes of targeting advertising. The specific formulation of the Pixel Privacy task addresses a highly simplified version of the overall problem of online multimedia privacy. The goal is to provide a foundation upon which solutions addressing progressively more realistic versions of the problem may be developed in the future.

The focus of the MediaEval 2018 Pixel Privacy task is on privacy-sensitive information that is related to scene categories. A scene category can be understood to be the identity of the setting in which a photo was taken. The task data is a subset of the Places365-Standard dataset [18]. The task provides a list of 60 privacy-sensitive categories chosen from the original 365 scene categories. It defines a validation set (MEPP18val) and a test set (MEPP18test) each containing 3000 images (50 from each of the 60 classes). Task participants use the validation set to develop their protection transformations. Then, they receive the test set, and are asked to apply their transformations to the test set images, and submit the protected test set images for evaluation.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

Although most protection transformations will be algorithms that are applied automatically by a computer, we also encourage participants (especially those specialized in art or photography) to develop manual protection techniques that are applied by hand. The purpose of manual techniques is to leverage creativity and explore unexpected new ways in which the visual appeal of images can be enhanced. Participants applying manual enhancement do not need to enhance 50 images for 60 categories. Instead, we have defined a special test set MEPP18test_manual, which is a subset of MEPP18test containing one image per category. Also, in case the user studies turn out to be too time-consuming, we will also focus the user study evaluation on MEPP18test_manual.

We inspect the scene categories and identify 60 scene categories that could be considered privacy-sensitive. Each class is related to each least one of ten privacy criteria: Places in the home, Places far away from the home (typical vacation places), Places typical for children, Places related to religion, Places related to people's health, Places related to alcohol consumption, Places in which people do not typically wear street clothes, Places related to people's living conditions/income, Places related to security, Places related to military. These privacy criteria are intended to represent aspects of privacy in images that are interesting for future work. The list provides a basis that can be refined or extended in the future.

3 EVALUATION

The submitted test set images will be evaluated with respect to *protection* and *appeal*. Performance of transformations with respect to *protection* is evaluated by measuring the degree to which the protected test set images block inference of a computer vision algorithm, referred to as the *attack algorithm*. The attack algorithm that we use is a ResNet50 [10] classifier trained on the training set of the Places365-Standard dataset [18]. 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.

We provide here some notes on the difference between the training data of the attack algorithm and the MEPP18val and MEPP18test datasets. MEPP18val and MEPP18test are derived from the *validation* set (and not the training set) of the Places365-Standard dataset, meaning that the attack algorithm is trained on data mutually exclusive from the Pixel Privacy validation and test sets. Note also that the Places365-Standard dataset was created by collecting images online, and for this reason, we do not expect any of the images in a given category to be different views of the same scene. We also assume that the data does not include any other forms of near duplicates.

Performance of transformations with respect to *appeal* will be carried out with respect to an automatic aesthetics classification algorithm, and also user study. NIMA [17] is used to evaluate the perception of the transformations aesthetics from the perspective of algorithms. The model is trained on AVA [16] dataset. Participant submissions will be ranked by the average difference in NIMA score and variances of the original and transformed image for the test set. Appeal will also be evaluated by a set of human annotators, who will inspect the original image and the enhanced image, and give a rating to the acceptability of the change. The ratings will be collected on a 7-point scale, according to whether the annotator agrees to the statement that the enhancement is more appealing for the scale points (1) strongly disagree (2) agree (3) somewhat agree (4) neither agree nor disagree (5) somewhat disagree (6) disagree (7) strongly disagree.

	Top-1 acc.	Top-5 acc.
Original images	59.60%	88.70%
Protected images	36.77%	64.07%
Protection gain (abs.)	22.83%	24.63%

Table 1: Top-1 and top-5 scene category predication accuracy on MEPP18val before and after protection by applying enhancement with CartoonGAN. The absolute difference is the protection gain.

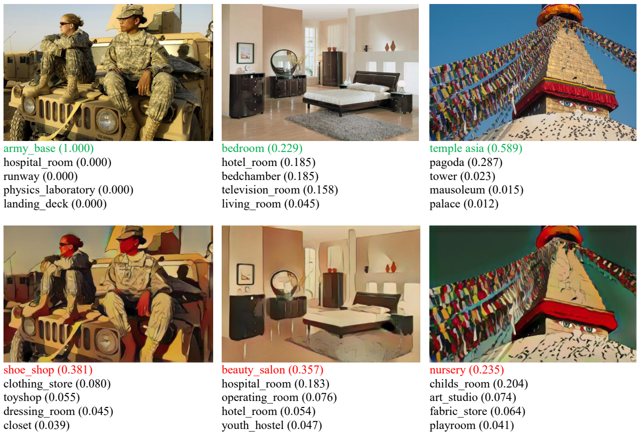

Figure 1: The predictions for three original images (top row) from MEPP18val and their enhanced versions (bottom row). For each image, the top-5 prediction labels and the corresponding confidences are given, where ground-truth labels are marked in green and the wrong top-1 prediction in red.

4 SIMPLE BASELINE

Here, we provide a simple baseline approach to enhancement by using a style transfer technique called CartoonGAN [4], based on a Generative Adversarial Network (GAN). We note that cartooning approaches to privacy in surveillance video have been previously used in the MediaEval Visual Privacy task [7, 8]. Table 1 shows the top-1 and top-5 predication accuracy before and after the enhancement. Figure 1 shows some image examples from three different categories (i.e., *army base*, *bedroom* and *temple asia*) from MEPP18val and their enhanced versions, along with the predicted results before and after the enhancement. In these examples, the image is correctly classified into a privacy-sensitive category before enhancement, but once the image is enhanced, the classifier can no longer predict the correct category. Inspection of these examples provides an impression of the potential user appeal of this form of privacy protection.

REFERENCES

[1] Atta Badii amd Tomas Piatrik, Mathieu Einig, and Chattun Lallah. 2012. Overview of MediaEval 2012 Visual Privacy Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop*. http://ceur-ws.org/Vol-927/mediaeval2012_submission_8.pdf

[2] Atta Badii, Pavel Koshunov, Hamid Oudi, Touradj Ebrahimi, Tomas Piatrik, Volker Eiselein, Natacha Ruchaud, Christian Fedorczak, Jean-Luc Dugelay, and Diego Fernandez Vazquez. 2015. Overview of the MediaEval 2015 Drone Protect Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop*. http://ceur-ws.org/Vol-1436/Paper7.pdf

[3] Serdar Çiftçi, Ahmet Oğuz Akyüz, and Touradj Ebrahimi. 2018. A Reliable and Reversible Image Privacy Protection Based on False Colors. *IEEE Transactions on Multimedia* 20, 1 (Jan 2018), 68–81.

[4] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. 2018. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9465–9474.

[5] Jaeyoung Choi, Martha Larson, Xinchao Li, Kevin Li, Gerald Friedland, and Alan Hanjalic. 2017. The Geo-Privacy Bonus of Popular Photo Enhancements. In *Proceedings of the 2017 ACM International Conference on Multimedia Retrieval (ICMR'17)*. 84–92.

[6] F. Dufaux and T. Ebrahimi. 2010. A framework for the validation of privacy protection solutions in video surveillance. In *2010 IEEE International Conference on Multimedia and Expo*. 66–71. https://doi.org/10.1109/ICME.2010.5583552

[7] Ádám Erdélyi, Thomas Winkler, and Bernhard Rinner. 2013. Serious Fun: Cartooning for Privacy Protection. *Working Notes Proceedings of the MediaEval 2013 Workshop*. http://ceur-ws.org/Vol-1043/mediaeval2013_submission_74.pdf

[8] Ádám Erdélyi, Thomas Winkler, and Bernhard Rinner. 2014. Multi-Level Cartooning for Context-Aware Privacy Protection in Visual Sensor Networks. In *Working Notes Proceedings of the MediaEval 2014 Workshop*. http://ceur-ws.org/Vol-1263/mediaeval2014_submission_41.pdf

[9] 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'10)*. 1–8.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 770–778.

[11] Pavel Korshunov and Touradj Ebrahimi. 2013. PEViD: privacy evaluation video dataset. *Proc. SPIE 8856, Applications of Digital Image Processing XXXVI*.

[12] 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 15-Sept-2018.

[13] McNamee, Roger and Parakilas, Sandy. 2018. The Facebook breach makes it clear: data must be regulated, The Guardian, 19 March. (2018). https://www.theguardian.com/commentisfree/2018/mar/19/facebook-data-cambridge-analytica-privacy-breach, Online; accessed 15-Sept-2018.

[14] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1765–1773.

[15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 2574–2582.

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

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

[18] 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* (2017).