Towards a Human in the Loop Approach to Preserve Privacy in Images

Andrea Mauri[1], Alessandro Bozzon[1]

[1]Delft University of Technology, Mekelweg 5, 2628 CD Delft

Abstract

Current artificial intelligence and information retrieval systems need to be trained with a large amount of data to achieve satisfying performance. A popular solution to create such datasets is to employ crowdsourcing; however, the content to be annotated may contain private or sensitive information that can be extracted by workers, limiting the applicability of crowdsourcing data annotation techniques in privacy-sensitive contexts. In this paper, we survey the literature finding that current solutions in crowdsourcing and machine learning do not provide satisfactory solutions as they either hinder the capabilities of workers to annotate the data, increase the overall cost, or lack generalizability. We identify current challenges, propose and elaborate a hybrid human-machine approach to detect private information in images, discuss its features and propose future directions.

Keywords

crowdsourcing, privacy preservation, human in the loop

1. Introduction

Crowdsourcing is currently employed for content annotation in various domains: generating ground-truth data for machine learning models, enrichment of multimedia data for better content retrieval and exploration, opinion mining, etc. Crowdsourcing tasks are typically performed by anonymous workers operating online microwork marketplaces (e.g., Amazon Mechanical Turk), or by contract workers operating on private platforms (e.g., oDesk, freelancer.com). A problem common to both types of platforms concerns potential exposure to workers of private information regarding the requester (i.e., the entity having the need for content annotation) or other people related to the content (e.g. owners of invoices to be digitized).

This is especially true for images; consider for instance a crowdsourcing task where workers are asked to transcribe the text contained in an image, such as in Figure 1. There, workers can access the information required to complete the task (i.e., the text on the banner), while being able to infer a wide range of private information: the identity of the people from their faces; their political alignment and, to some extent, where they live from the sign. A real-world example is the case of the security cameras sold by the Ring startup. It was recently discovered that to train and improve the performance of their machine learning model, the employees had total and unfiltered access to the live feed of the cameras[1].

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

a.mauri@tudelft.nl (A. Mauri); a.bozzon@tudelft.nl (A. Bozzon)

0000-0002-1263-4575 (A. Mauri); 0000-0002-3300-2913 (A. Bozzon)

[1]https://techcrunch.com/2019/01/10/amazon-ring-privacy-concerns/

Figure 1: Example of image containing private information [1]

While current research on privacy preservation in crowdsourcing focuses on worker privacy [2], by developing user modeling and task assignment methods that exploit user properties while protecting their private information, less attention has been devoted to the issues of content privatization - i.e, developing methods that reduce the risk of leaking private information from the content of the task, while still allowing the workers to complete it. Given the growing need for annotated data, the demand for privacy-aware content management, and stricter privacy protection regulation, there is a clear need for efficient and effective methods for privacy preservation in the content of crowdsourcing tasks.

Automated methods for privacy preservation can help in terms of efficiency; however, their utility is limited by the quality of their output, by the harmful mistakes they might make in the privacy context, and by their need for a large amount of training data that forces to trade-off between cost and generalization abilities. This is particularly true for images, where works focus on obfuscating only particular instances of private data (e.g., faces) [3, 4, 5] by automatic detection. To the best of our knowledge, only Orekondi et al. [1] attempted to use machine learning to both detect and obfuscate private information in images.

Works in the crowdsourcing domain deal with privacy by either obfuscating entirely the content of the tasks [6, 7, 8], making it more difficult for the workers to complete their work, or requiring input from the requester [9], suffering from scalability problem. Others segment the image in small patches [10, 11], so a worker does not see the whole picture. While having a minor impact on the performance of the workers, they cost more, since every single patch needs to be examined.

In this paper, we first contribute with a survey on the state-of-the-art methods to protect

privacy in the content shown in crowdsourcing tasks. We investigate how current work in crowdsourcing deals with the private information contained in the task content, and we explore which automatic methods could be used to protect privacy in crowdsourcing tasks.

We elaborate on the strengths and weaknesses of different classes of approaches, and we propose a design of a human-machine pipeline to detect and obfuscate private content in images more effectively and efficiently than existing methods. We envision a combination of pre-trained computer vision models for understanding image content in order to reduce the cost of developing the system (i.e., no need for developing a new training dataset and training a new deep learning model) and capabilities of visual and logical reasoning of humans.

2. Literature Review

We adopted a Systematic Literature Review approach [12], searching papers through Google Scholar.

We strove for a balance between crowdsourcing and machine learning approaches. On average, we explored 30 result pages for searches including keywords "privacy" and ("preservation" or "preserving") and "crowdsourcing" and ("guarantees" or "tasks" or "surveys"). Similarly, works related to privacy preservation in machine learning are searched with the keyword combination of "privacy" and "preservation" and "machine learning" and ("surveys" or "images" or "personal information").

Using a snowball strategy, we explored references contained in the retrieved papers. We disregard works addressing the problem of preserving the privacy of workers in crowdsourcing. In the search of papers on machine learning approaches, we exclude methods focusing on machine-to-machine privacy preservation (i.e., privacy-aware version of existing algorithms) and works that images completely unintelligible.

We classify the resulting papers according to the adopted methods for private information detection, and techniques for privacy preservation.

2.1. Methods to preserve content privacy in crowdsourcing

Crowdsourcing works that deal with privacy preservation attempt to make the crowdsourcing workflow privacy-aware by either obfuscating the content of the task (e.g. by applying filters such as blur and visual noise), or segmenting it in small patches in such a way that workers never see the whole content.

With images, the majority of works do not focus on information detection, but on information preservation, by studying how different types of obfuscation methods influence the quality of the responses of the crowd.

Varshney et al., [7] use a combination of random noise and error correction codes [13] to mask private information in images and studied the trade-off between the amount of privacy preserved and the utility of the picture (i.e., to what extent the image can be still used by the crowd). They notice that by obfuscating the content of the task, the quality of the answers decreases, but adding the error correction code greatly improves the quality of the responses.

In [8] authors explore how much blurring effects on video frames influence the accuracy of the response received in an image annotation task. Similar to [7], the objective is to study

the trade-off between preserving privacy and the accuracy of the responses. The significant difference between the two papers is in the measurement of the trade-off between accuracy and privacy, which is done respectively through experimentation and theoretical proof. They conclude that increasing the level of blurring resulted in a performance degradation similar to the estimates of [7], suggesting data obfuscation method needs will give a balance between the reliability of responses from the workers and the level of privacy preserved on the task content. Alshaibani et al. [14] use a similar method, but it iteratively reduces the amount of obfuscation and asks a group of workers to review the image at each step. Kajino et al. [10] propose a different strategy, consisting in segmenting the task content in small clips. The intuition is that by segmenting the content (e.g., images, text, or audio), the probability a worker sees sensitive or private information decreases.

Works focusing on the detection of private information, use inputs either from the requester or the crowd itself. In Zensors++ [9], the system automatically obfuscates people's faces using a face detection model and the requester selects the region of the image that needs to be sent to the workers.

CrowdMask [11] leverages the crowd to identify and obfuscate private information in images. The main idea is to segment and distribute small patches of the image to the workers so they can detect possible private information without seeing the whole picture. They propose an iterative workflow; at every step the image is segmented and the segments are sent to the crowd to be annotated. Segments containing private information are obfuscated, and the process is then repeated using bigger patches.

2.2. Automatic and Machine Learning methods to preserve privacy

In this section, we review automatic and machine learning methods that can be used to detect and preserve privacy in the content used in crowdsourcing tasks.

We can identify two main categories of research: studies that focus on building machine learning models to *detect* private information and works dedicated to create methods to *obfuscate* the private information while not disrupting the utility of the image.

An attempt to create a general privacy detection model was done by Orekondi et al. [1]. They use an ensemble of machine learning models able to recognize and mask the private information present in the image. To the best of our knowledge, this is the first work that extends an object detection approach to detect private information in images.

Recently, another dataset was released by Gurari et al. [15] The pictures were obtained from their VizWiz real-time visual question answering system [16] , and they were manually annotated by the authors using a 23 categories taxonomy.

Uittenbogaard et al. [17] developed a method to detect, remove and inpaint private visual cues in street-level imagery (e.g., people and vehicles). Assuming that in this kind of image moving objects are the ones at risk of disclosing private information, they develop a method that detects moving objects by exploiting inconsistencies between different frames.

Work developing new obfuscation methods focuses on specific instances of private information (e.g., faces). Generative adversarial networks have been used to anonymize faces by removing privacy-sensitive information [5], replace them with cartoons [4] or generating fake ones [18]. By doing so, they protect the identity of the person appearing in the image while keep-

ing the facial expression. Chhabra et al. [3] developed a k-anonymity algorithm to anonymize selective attributes in a face, without compromising the quality of the picture.

2.3. Discussion and challenges

Currently, works in crowdsourcing deal with privacy by either obfuscating entirely the content of the tasks, making it more difficult for the workers to complete their work, or requiring input from the requester, suffering from scalability problems. Approaches based on segmentation seem promising since they have a minor impact on the performance of the workers. On the other hand, they cost more, since workers need to examine patches and not the whole content. Most of the automatic approaches focus on blocking private information disregarding its detection. For this reason, they are tied to specific use cases (e.g., obfuscating faces). To the best of our knowledge, a general-purpose model for privacy preservation is still missing. The model presented in [1] shows inconsistent performances between different classes and the categories do not cover all the types of private information.

We identify two main challenges toward an effective detection of private information:

1. The lack of a clear understanding of what private information is. While this is obvious for text, connecting the visual cues present in an image with certain classes of private information is more difficult. A face can be straightforwardly connected to a person's identity, but what about the clothes someone is wearing? Or the items present in a room? As stated in the *General Data Protection Regulation* (GDPR)[2], not only information directly identifying a person should be protected, but also any clue about sensitive or personal information. Defining privacy is a complex problem, as it requires translating the concepts described in policies and laws into entities and relations understandable by machines.
2. The lack of a comprehensive dataset to train a general-purpose machine learning model. The datasets reviewed in this paper [1, 15] were manually built by the authors, requiring an enormous amount of time (the authors of [1] claim it took them 800 hours over 4 months to annotate the dataset). We cannot use directly crowdsourcing to scale the dataset annotation, because we cannot risk showing private information to the crowd. We need to design a crowdsourcing process that can mask the private information while still allowing the crowd to annotate it, that is the main challenge that motivated this paper. Segmentation-based approaches seem promising, but, as stated before, they come with an increased cost.

Concluding, state-of-the-art shows privacy preservation in the context of crowdsourcing tasks cannot be solved effectively and efficiently by the crowd or machine learning alone. We argue that by combining the flexibility of the crowd with the cost-effectiveness of automatic approaches it is possible to improve the performance with respect to the state of the art. Hybrid human-machines solutions have been successful in many fields [19], but, to the best of our knowledge, their application for privacy preservation is still unexplored.

[2] https://gdpr-info.eu/

3. Proposed approach

3.1. Clarifying the concept of private information

To understand what private information is, and to provide an unbiased definition, we look at definitions present in the *General Data Protection Regulation* (GDPR)[3], a legal act dictating how personal data needs to be protected during its collection and processing. In this policy personal data is defined as follow:

> (personal data) is any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person

Here we can identify private information (e.g., age, face, address, etc..), sensitive information (e.g, sexual orientation, trade or political memberships), or cues that can be used to infer personal information (e.g, economic, physiological, social factors like lifestyle, etc).

While this is a step forward in framing the concept of private information, it is not precise enough in the case of images. The identity of a person can easily be connected to his/her face, but what about his or her economic, cultural or social status?

To get a better understanding we use the Data Privacy Vocabulary [20] drafted by the Data Privacy Vocabularies and Controls Community Group. The vocabulary provides classes and properties describing instances of private data, the process they undergo, and who manages them. In this work, we consider only the categories of data (i.e., the *Personal Data Category* class) and we limit the scope to the ones that can reasonably appear in images. We exclude classes referring to audio (e.g., *Voice Communication Recording*), biological (e.g., *DNACode*) and device-based tracking data (e.g., *IPAddress*).

Private data is divided in six main categories: *Internal*, *External*, *Financial*, *Social*, *Tracking* and *Historical*. *Internal* comprehends all the information about an individual that - usually - cannot be observed because they are either kept secret (i.e., *Authenticating* data such as passwords or pincode) or they are part of someone set of believes (e.g., *Religous Belief*, *Opinion*, etc..).

External includes all the characteristics of a person that can be from his/her physical appearance such as *Demographic*, *PhysicalCharacteristic* and *Ethnicity*. *Financial* refers to information about monetary transactions and ownership. *Social* include data about social aspects of an individual like family, public life, or professional networks. *Tracking* refers to the information that can be used to detect the location of an individual. Finally, *Historical* refers to the historical events a person witnessed.

3.2. Hybrid human-machine approach to detect and obfuscate private information

In this section, we propose a hybrid human-machine approach to detect and obfuscate private information in images, where machine learning models inform the process of involving human

[3] https://gdpr-info.eu/

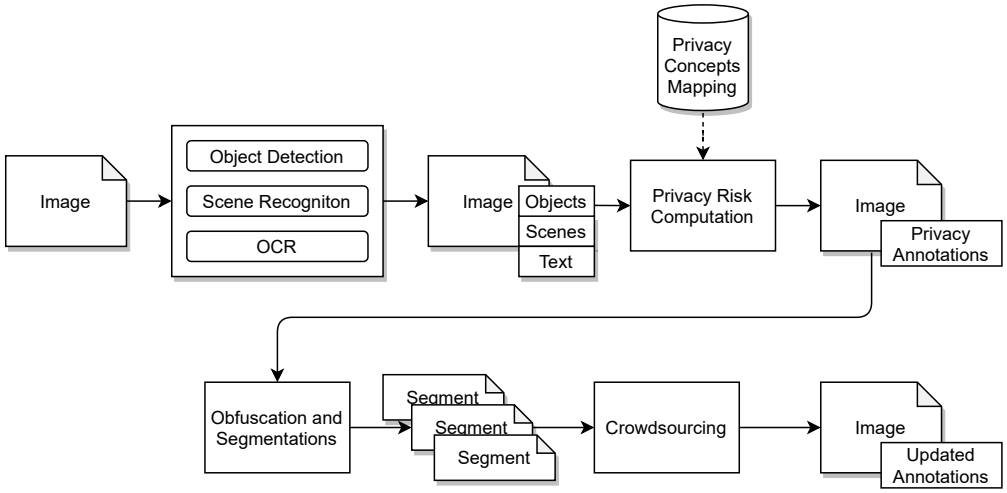

Figure 2: Overview of the envisioned system

intelligence with the generation of crowdsourcing tasks.

Previous works to detect private information in images have shown that machine learning approaches are costly and non-robust since they require building large training datasets, and the trained models can suffer from non-generalization issues to other datasets due to biases. That is why we choose to leverage pre-trained machine learning models as they do not add any cost to the pipeline, while there are many diverse ones available publicly.

Our intuition is that even out-of-the-box machine learning models can provide useful information about the private content included in an image, even though their outputs do not directly map to private information.

For instance, an object detection model such as Mask R-CNN [21] can provide labels as `person` - that can be easily mapped to the entities described in the previous section - or `laptop` that can indicate the presence of *Authenticating* of *Financial* information, as they may appear on the screen. Also a scene detection model [22] can provide useful insight. For example, a scene detected as `indoor` may contain information about the work environment, confidential documents, and computer screens.

In addition, work in the crowdsourcing domain show that approaches based on segmentation seems to be effective in hiding private information from the workers, with the negative aspect, however, of an increased cost. Here our idea is that the annotations provided by the machine learning models can be used to reduce the number of patches the crowd needs to analyze by either find part of the image that does not contain any element, or, on the contrary, flagging sections of the images where there are visual cues of private information (e.g, a face).

For these reasons, we envision our pipeline - shown in Figure 2 composed by these steps. First, an ensemble of pre-trained machine learning models for object detection, scene detection, and optical character recognition is used to detect the different visual cues in the image. The output consists of: object labels, object masks/bounding boxes, texts in the image, scene attributes, and scene categories along with the confidence scores of the detected objects and scene annotations. The idea is that the label produced by the machine learning models can guide the segmentation

and the crowdsourcing phase by providing a machine-readable description of the scene.

For example, let us consider an image of a workplace that has objects like a computer, and few documents on the table - the associate scene attributes for this image from scene detection will be man-made, paper, working, studying, research, etc. give context to the image based on which we perform segmentation. The annotation is mapped to the private entities described in Section 3.1.

The result of this mapping, together with the confidence score of the machine learning models, is used to compute a *privacy leak risk* score of the image. The scene detection model contributes to a global score, as - intuitively - photos of some places are more at risk of showing sensitive information than others (e.g, an office versus a landscape). If the local score (combined with the global) is above a threshold, the object mask is obfuscated.

Then, the image is segmented in small patches, and only the segments that are not completely obscured - and have a privacy leak score falling in a given range - are sent to the crowd for further evaluation. Similar to previous works [11, 10], the idea is that by showing only a limited portion of the image, it is possible to detect private information without revealing sensitive content to the workers.

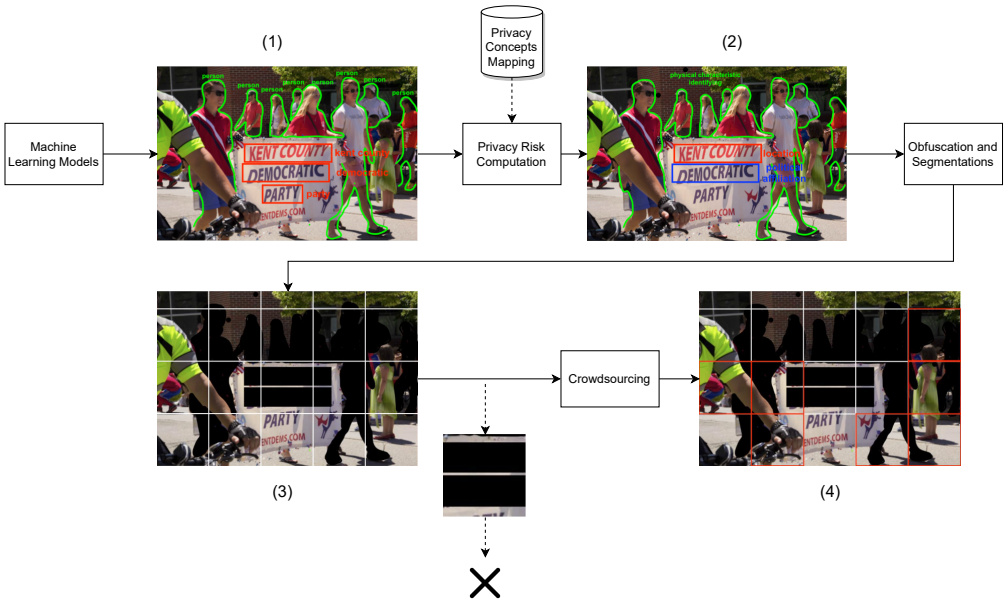

Figure 3: A running example of the pipeline

To better illustrate our approach, Figure 3 shows a running example of the pipeline. First, the machine learning models[4] detect people in the picture and extract the text present in the banner (1). Then, we use the mapping with the privacy concepts to identify where in the image they appear (2). If the algorithm is confident enough, it obfuscates the part of the image showing those concepts, then it segments it in small patches for the crowd for further evaluation. Notice

[4]To keep it simple we only consider an object detection model and OCR

that patches that are fully obscured are not sent (3). Finally, the crowd further analyzes the segments and flags the one containing private information (4).

Each step of the pipeline has hyperparameters on which to experiment in order to analyze the effectiveness and efficiency of the component. Such hyperparameters are the following.

Choice of the machine learning models Different machine learning models allow us to get different information from an image. Object detection algorithms can find relevant visual elements present in the image, scene detection models can give an additional indication on the setting shown in the picture, while OCR can extract possible sensible or confidential information from the text.

Mapping the labels with the private entities This mapping can be created in different ways: it can be manually built considering the machine learning model used - i.e., hard-coded -, or automatically defined using, for instance, a rule-based approach - i.e., the logic combination of terms correspond to a private term -, a probabilistic method - i.e., a term has a probability to refer to a private attribute - or learned - i.e., the probability of the terms are learned in an iteratively way during the process.

Threshold for the privacy risk score This parameter regulates when a part of an image needs to be obfuscated, sent to the crowd, or be considered safe. A high threshold will lead to parts of the image be wrongly assumed as safe or with many segments sent to the crowd for further analysis; while a value too low will make the pipeline preventively obscure the majority of the image.

Design of the crowdsourcing task The design of the crowdsourcing task can greatly influence the outcome of the whole pipeline. The amount incentives, the design of the interface, including the type of task the crowd needs to perform impact the quality of the results and define the type of annotations provided.

4. Conclusion and Future Works

In this paper, we argue the task of privacy preservation in images cannot be solved effectively and efficiently by the crowd or machine learning alone. We surveyed the state of the art both in crowdsourcing and machine learning domains and found out the main challenges related to privacy preservation lie in the complexity of understanding what privacy is, and in the difficulty of creating a dataset to train machine learning models.

To address these challenges, we then propose a hybrid human-machine pipeline to detect private information in images. The pipeline utilizes pre-trained machine learning models to detect visual cues of private information in images and uses crowdsourcing to detect private information the machine learning models may have missed. In this way, it harnesses the strength of both approaches by combining the efficiency of automatic methods and the flexibility of human intelligence.

Future works will focus on the implementation and deployment of the before-mentioned pipeline to validate the approach and refine, through detailed experimentation, the various components. We will investigate how different machine learning models and crowdsourcing task designs influence the outcome of the pipeline. Finally, we will evaluate the impact of privacy preservation on the performances of artificial intelligence and information retrieval system.

References

[1] T. Orekondy, M. Fritz, B. Schiele, Connecting pixels to privacy and utility: Automatic redaction of private information in images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8466–8475.

[2] F. Khan, A. U. Rehman, J. Zheng, M. A. Jan, M. Alam, Mobile crowdsensing: A survey on privacy-preservation, task management, assignment models, and incentives mechanisms, Future Generation Computer Systems 100 (2019) 456–472.

[3] S. Chhabra, R. Singh, M. Vatsa, G. Gupta, Anonymizing k-facial attributes via adversarial perturbations, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18, AAAI Press, 2018, pp. 656–662. URL: http://dl.acm.org/citation.cfm?id=3304415.3304509.

[4] J. Chen, J. Konrad, P. Ishwar, Vgan-based image representation learning for privacy-preserving facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1570–1579.

[5] H. Hukkelås, R. Mester, F. Lindseth, Deepprivacy: A generative adversarial network for face anonymization, in: International Symposium on Visual Computing, Springer, 2019, pp. 565–578.

[6] L. R. Varshney, Privacy and reliability in crowdsourcing service delivery, in: 2012 Annual SRII Global Conference, 2012, pp. 55–60. doi:10.1109/SRII.2012.17.

[7] L. R. Varshney, A. Vempaty, P. K. Varshney, Assuring privacy and reliability in crowdsourcing with coding, in: 2014 Information Theory and Applications Workshop (ITA), 2014, pp. 1–6. doi:10.1109/ITA.2014.6804213.

[8] W. S. Lasecki, M. Gordon, W. Leung, E. Lim, J. P. Bigham, S. P. Dow, Exploring privacy and accuracy trade-offs in crowdsourced behavioral video coding, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, ACM, New York, NY, USA, 2015, pp. 1945–1954. URL: http://doi.acm.org/10.1145/2702123.2702605. doi:10.1145/2702123.2702605.

[9] A. Guo, A. Jain, S. Ghose, G. Laput, C. Harrison, J. P. Bigham, Crowd-ai camera sensing in the real world, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2 (2018) 111.

[10] H. Kajino, Y. Baba, H. Kashima, Instance-privacy preserving crowdsourcing, in: Second AAAI Conference on Human Computation and Crowdsourcing, 2014.

[11] H. Kaur, M. Gordon, Y. Yang, J. P. Bigham, J. Teevan, E. Kamar, W. S. Lasecki, Crowdmask: Using crowds to preserve privacy in crowd-powered systems via progressive filtering, in: Fifth AAAI Conference on Human Computation and Crowdsourcing, 2017.

[12] D. Gough, S. Oliver, J. Thomas, An Introduction to Systematic Reviews, SAGE Publications, 2017. URL: https://books.google.nl/books?id=41sCDgAAQBAJ.

[13] T.-Y. Wang, Y. Han, P. Varshney, P.-N. Chen, Distributed fault-tolerant classification in wireless sensor networks, IEEE Journal on Selected Areas in Communications 23 (2005) 724–734. doi:10.1109/JSAC.2005.843541.

[14] A. Alshaibani, S. Carrell, L.-H. Tseng, J. Shin, A. Quinn, Privacy-preserving face redaction using crowdsourcing, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (2020) 13–22. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7459.

[15] D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, J. P. Bigham, Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 939–948.

[16] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al., Vizwiz: nearly real-time answers to visual questions, in: Proceedings of the 23nd annual ACM symposium on User interface software and technology, ACM, 2010, pp. 333–342.

[17] R. Uittenbogaard, C. Sebastian, J. Vijverberg, B. Boom, D. M. Gavrila, et al., Privacy protection in street-view panoramas using depth and multi-view imagery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10581–10590.

[18] Z. Chen, T. Zhu, C. Wang, W. Ren, P. Xiong, Gan-based image privacy preservation: Balancing privacy and utility, in: X. Chen, H. Yan, Q. Yan, X. Zhang (Eds.), Machine Learning for Cyber Security, Springer International Publishing, Cham, 2020, pp. 287–296.

[19] J. W. Vaughan, Making better use of the crowd: How crowdsourcing can advance machine learning research., Journal of Machine Learning Research 18 (2017) 193–1.

[20] Data Privacy Vocabularies and Controls Community Group, Data privacy vocabulary v0.1, 2019. https://www.w3.org/ns/dpv, Last accessed on 2021-07-01.

[21] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[22] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 1452–1464.