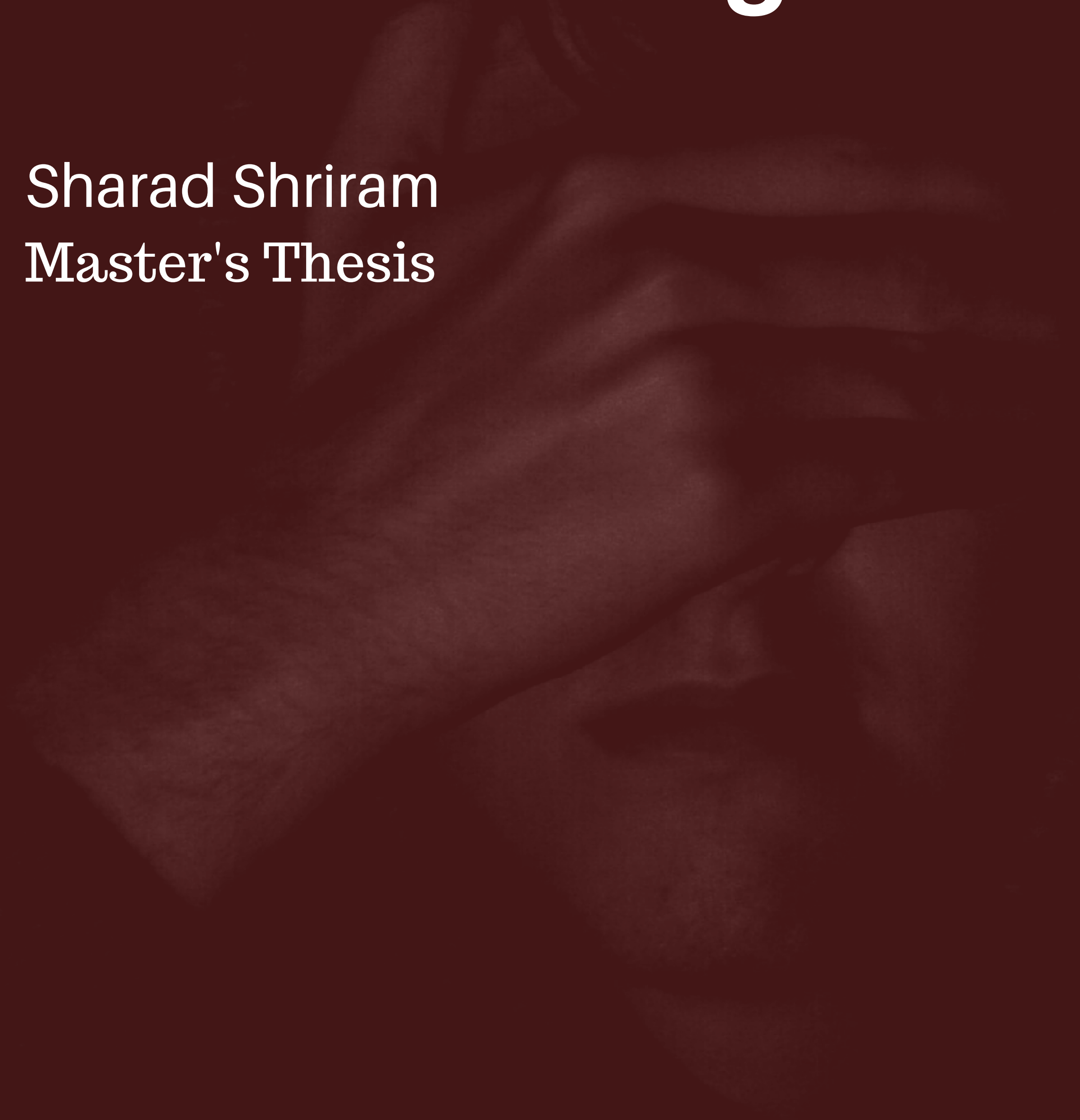
A Human-Machine Approach to Preserve Privacy in Image Analysis Crowdsourcing Tasks



Sharad Shriram

Master's Thesis

TU Delft

A Human-Machine Approach to Preserve Privacy in Image Analysis Crowdsourcing Tasks

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK DATA SCIENCE AND TECHNOLOGY

by

Sharad Shriram
born in Mumbai, India

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

A Human-Machine Approach to Preserve Privacy in Image Analysis Crowdsourcing Tasks

Author:	Sharad Shriram
Student id:	4671082
Email:	S.Shriram@student.tudelft.nl

Abstract

Modern web information systems use machine learning models to provide personalized user services and experiences. However, machine learning models require annotated data for training, and creating annotated data is done through crowdsourcing tasks. The content used in annotation crowdsourcing tasks like medical records and images might contain some private information which can directly or indirectly identify an individual. The name, age, ethnicity, gender, contact details are examples of private information that directly identifies an individual. Indirect private information relates to the cultural, economic, and social factors of an individual. For instance, the visual cues of religious objects or symbols relate to the religious beliefs of an individual. In this thesis, we study how to minimize the amount of private information extracted from images using a hybrid algorithm which combines machine learning models and crowdsourcing. We also demonstrate that the proposed hybrid algorithm reduces the amount of private information exposed from the image and the cost of using the crowd for detecting private information in the image.

Thesis Committee:

Chair:	Prof. dr. ir. G.J.P. M. Houben, Faculty EEMCS, TUDelft
University supervisor:	Prof. dr. ir. A. Bozzon, Faculty EEMCS and IO, TUDelft
Daily supervisor:	Dr. A. Mauri, Faculty EEMCS, TUDelft
Committee Member:	Dr. ir. M. Finavaro Aniche, Faculty EEMCS, TUDelft

Preface

This thesis titled “A Human-Machine Approach to Preserve Privacy in Image Analysis Crowdsourcing Tasks” is done in partial fulfillment of the requirements for the MSc in Computer Science degree at the Delft University of Technology. The work on this thesis spanned from the period of 15 September 2018 to 19 August 2019.

Through this thesis, we contribute a hybrid human-machine approach to preserve privacy in images used in image analysis crowdsourcing tasks. With policies like the GDPR, data-driven workflows must become privacy-aware. Privacy preservation in image analysis tasks hinges on the trade-off between the cost to preserve privacy and the usefulness of the privacy-preserved image in image analysis tasks. Our proposed approach balances this trade-off by combining human and machine intelligence, which makes the proposed approach effective in detecting private information through visual cues in images.

The thesis started as a flowchart which Prof. Alessandro Bozzon, my supervisor drew on a whiteboard in our first meeting. I am grateful to Prof. Bozzon for the opportunity to work on this topic and his feedbacks during crucial stages of the thesis. The evolution of this topic from the whiteboard to a prototype had a generous amount of hurdles, making it challenging and fun to work. At this point, I am grateful to Dr. Andrea Mauri, my daily supervisor, for making time to guide and support me clearing the many hurdles during this thesis. He was always there to listen patiently to my many brainstorms and would later put me back on track to work on completing the thesis. I am grateful to the committee members, Prof. Geert-Jan Houben, and Dr. Mauricio Aniche for their time, availability and feedback on the thesis.

I am grateful to my parents and my friends for their encouragement and support.

Sharad Shriram
Delft, the Netherlands
August 12, 2019

Contents

List of Figures

Chapter 1

Introduction

Modern web information systems use machine learning models to provide a personalized experience for user convenience. However, these machine learning models need an annotated dataset to learn and identify features to improve user convenience. Annotating datasets is typically done by humans through crowdsourcing. Annotating image datasets through crowdsourcing falls under the category of image analysis crowdsourcing tasks.

Image annotation through crowdsourcing is done by first distributing the image to a pool of remote people referred to as “the crowd” or crowd workers. Based on the requirements and instructions given for the task, the crowd responds and completes the task. In image annotation tasks, the crowd generally respond by marking regions on the image or through textual descriptions. While responding to annotation tasks, the crowd can extract additional information from the image which may be irrelevant to complete the task. The extracted information by the crowd can contain private information like the name, contact details, religious beliefs, political views, social life, and economic well-being of an individual. With policies like the GDPR, it is a must to preserve the privacy of private information[1] in workflows involving data collection, processing, and storage.

For the images used in image analysis crowdsourcing tasks, privacy can be preserved using obfuscation methods like blurring [23], blocking [31] or by adding random distortions [42] on the private visual cues like the people in the image. However, there may be other visual cues containing private information which needs to be first detected and then obfuscated. We can broadly classify the current approaches for privacy preservation in images into machine learning-based or crowdsourcing-based approaches. As we show later in the thesis, these approaches suffer from the following limitations:

- Recent *machine learning-based approaches* are showing a notable increase in the accuracy of detecting private information in images. However, the accuracy of detecting private machine information using machine learning-based approaches depends on a dataset annotated for visual cues containing private information.

[1] https://gdpr-info.eu/art-4-gdpr/

- *Crowdsourcing-based approaches* for privacy preservation can yield fine-grained categorization of the private information present in the image. These approaches have a better accuracy over machine learning-based approaches for detecting private information in the image. However, scaling crowdsourcing-based approaches for detecting private information for a large collection of images makes it very expensive.

Crowdsourcing-based approaches are not only expensive for detecting private information for a large collection of images but also have the privacy concern of the crowd being able to extract private information from the images used in the crowdsourcing task [25]. Recent works have proposed task content segmentation [18, 20], task assignment methods [4] to reduce the amount of private information extracted by the crowd. However, these approaches are not cost-efficient when used for a large collection of images. The limitations of crowdsourcing-based approaches can be overcome using machine learning-based approaches like [31]. However, the number of private information detected using machine leaning-based approaches are limited to the annotated classes in the dataset used for training these models. Creating annotated datasets are typically done through crowdsourcing and the privacy concerns associated with crowdsourcing leads to a vicious loop between machine learning and crowdsourcing. The limitations of both machine learning-based and crowdsourcing-based approaches for detecting private information leads to our hypothesis that *combining machine learning and crowdsourcing approaches can result in a cost-efficient approach for detecting private information in images*.

1.1 Problem Statement

An example of a crowdsourcing image annotation task where a crowd worker is asked to annotate the image in Figure 1.1 for the scene category.

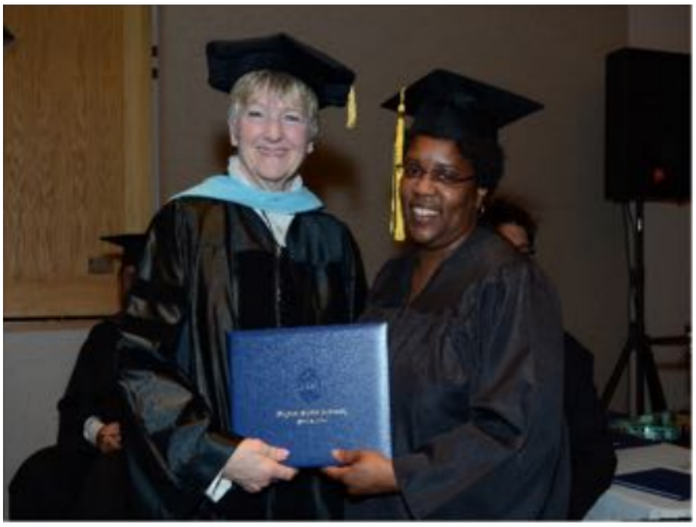


Figure 1.1: The extracted private in this image include the gender, ethnicity, profession of the people and the location [32].

Although the task asks for the annotation of the scene category, the crowd workers can extract additional information from the image before annotating the scene category as “graduation”. In this case, the extracted private information includes the location, gender, ethnicity of the people in the image. For this image, machine learning-based privacy preservation approaches will obfuscate the people in the image. However, obfuscating the people in the image makes it difficult for the crowd to annotate the scene category. In this case, preserving the privacy of the image using crowdsourcing-based approaches will be effective to collect the scene category annotations for the image. However, there is the privacy concern of the crowd workers being able to extract some private information from the image. Thus, we find the current approaches for privacy preservation in images requires us to do a trade-off between the amount of privacy preserved, the cost and usefulness of the image for further processing.

In this thesis, we study if we can leverage the advantages of both machine learning and crowdsourcing for detecting private information in images. We take inspiration from recent works [21, 19], which used the hybrid approach of combining crowdsourcing and machine learning for solving problems in different domains. However, these approaches are not suitable for detecting private information in an image, since they primarily were tested on textual data and there are no prior works which have mapped the different visual cues in an image. We aim to develop a hybrid algorithm which balances the amount of machine learning labels and crowdsourcing tasks used to detect private information in images. The objective of our hybrid algorithm would be to maximize the amount of private information detected and minimize the cost of using the crowd for detecting private information in the image. Thus, our main research question (MRQ) is as follows:

MRQ: How machine learning and crowdsourcing can be combined to efficiently and effectively detect and obfuscate private information in images used for image analysis crowdsourcing tasks?

The main research question is divided into three research sub-questions (RSQ):

RSQ 1: What are the current state of the art methods to detect and obfuscate private information through machine learning and crowdsourcing?

Recent works on privacy preservation through machine learning and in crowdsourcing can be classified based on the content used (image, or text), the approach to detect private information and the method to obfuscate the detected private information. These works typically study either the usefulness of the content after obfuscating the private information or the effectiveness of the approach to detect private information. The approaches to detect private information aims to identify descriptive features corresponding to private information in the content, like visual cues in images. The methods for obfuscating the detected private information is applicable only for machine learning-based works which study the effectiveness of privacy preserved using a machine learning model with different obfuscation methods. In crowdsourcing-based research, the focus is to measure the efficiency of the crowd for tasks using privacy-preserved content.

RSQ 2: How to combine machine learning models with crowdsourcing to detect private information in images?

There is limited availability of annotated image datasets for different private information and their corresponding visual cues. Crowdsourcing tasks to annotate private information to their corresponding visual cues can overcome the limited availability of annotated image datasets. However, state of the art approaches for preserving privacy in crowdsourcing are expensive at scale. Hypothetically, it should be possible to use pre-trained machine learning models as off-the-shelf components to detect and obfuscate some visual cues containing private information like the people in an image. The preliminary obfuscation through the machine learning model should reduce the number of private information visible to the crowd when asked to detect private information in the image or image segment [20]. Thus, we develop a hybrid algorithm which uses the predicted labels from the machine learning models to estimate the amount of private information likely to be disclosed in the image and determines the number of image segments used in crowdsourcing tasks to detect private information in the image.

RSQ 3: How to maximize the privacy preserved in the image while minimizing the cost of using the crowd for detecting private information in the image?

We study how the computed privacy disclosure for the image affects the number of crowdsourcing tasks created for detecting the private information in the image. To compute the privacy disclosure for the image, we count the number of detected visual cues containing private information in the image. Based on different thresholds of the computed privacy disclosure for the image, we study the number of crowdsourcing tasks created for detecting the private information in the image and also the size of the image segments used in these tasks. Since our objective is to use the privacy-preserved image for image analysis tasks, we need the qualitative evaluation of the usefulness of the privacy-preserved image for image analysis tasks and the amount of obfuscation used for preserving the privacy in the image.

1.2 Thesis Contributions

The original contributions (C) of this thesis based on the research sub-questions (RSQ) are as follows:

- **C1**: a comprehensive, systematic literature study on the current state of the art privacy preservation approaches for images through machine learning and in crowdsourcing, to answer RSQ 1.

- **C2**: a mapping of different categories of private information based on the GDPR to their corresponding visual cues in an image. This contribution branches from the answers to RSQ 1 and RSQ 2.

- **C3**: a hybrid algorithm which creates crowdsourcing tasks for detecting private information by segmenting the image based on the scene context and privacy disclosure computed using labels generated from machine learning models, contributing to RSQ 2.

- **C4**: a prototype to demonstrate the detection and obfuscation of private information in an image using our proposed hybrid algorithm, partly contributing to RSQ 3.
- **C5**: a quantitative study on the cost of using the crowd for detecting private information along with a qualitative study on the usefulness of the privacy preserved image together contributes to RSQ 3.

1.3 Research Outline

The organization of the remaining chapters of this thesis is as follows:

- Chapter 2 discusses the findings from the systematic literature review on privacy preservation through machine learning and crowdsourcing, identification of the existing research gap.
- Chapter 3 discusses in detail the design and definition of the proposed hybrid algorithm for detecting private information using the mapping of private information to their corresponding visual cues in an image based on the scene context.
- Chapter 4 presents the choices made during different stages of implementing the prototype to demonstrate the detection and obfuscation of private information in an image using our proposed hybrid algorithm.
- Chapter 5 discusses the experimental design and results of the quantitative study on the cost of using the crowd for detecting private information and the qualitative study on the usefulness of the privacy preserved image for image analysis tasks.
- Chapter 6 is the conclusion of this thesis report, where we discuss in brief the contributions by revisiting the research questions and also share possible research directions for future work.

Chapter 2

Background and Related Work

In this chapter, we discuss the current state of the art privacy preservation approaches for images through machine learning and in crowdsourcing, resulting in the thesis contribution C1.

In the past decade, the topics of detecting private information and preserving privacy have attracted contributions from different communities like computer vision, deep learning, and crowdsourcing. Recently, any study on privacy preservation features the widely used and popular privacy preservation approaches of differential privacy, k-anonymity, and t-closeness. Since we focus on privacy preservation approaches for images, the only related work using differential privacy for images was by Fan [11] to preserve the privacy of license plate and entities by pixelating the image. In this study, we do not consider semantic segmentation based privacy preservation approaches for street-level images since they are use-case specific and has a finite category of private visual cues. To the best of our knowledge, Orekondy et al. [31] is a representative of the effectiveness of using machine learning models to detect and obfuscate private information in images.

The topic of privacy preservation in crowdsourcing is classified based on the type of privacy preserved as worker-based and task content-based methods. Worker-based privacy preservation methods preserve private information related to the identity, location of workers participating in crowdsensing [44] or participatory crowdsourcing [17] tasks. Task content-based privacy preservation methods aim to limit the number of private information visible in the task content through content segmentation approaches [20], privacy-aware task assignment strategies [4] and task content obfuscation through random perturbations [41, 42]. In this study, we will focus on task content-based privacy preservation methods in crowdsourcing.

Strategy for finding literature

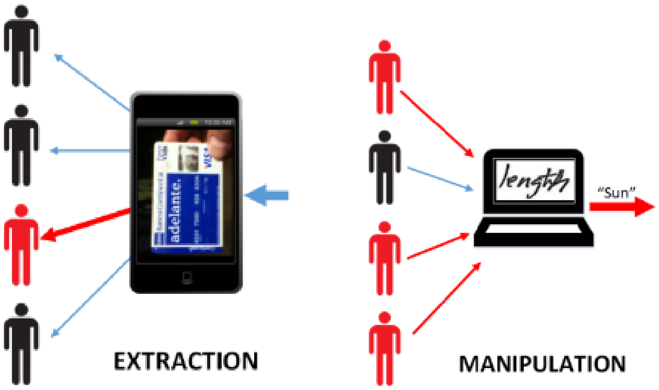
The research works referred to in this study are searched through Google Scholar. While selecting literature for the study, we focused on optimizing our search for the most recent and relevant research in both machine learning and crowdsourcing. The following keyword logic was used to collect most of the literature included in this study. For crowdsourcing we use,“privacy” and (“preservation” or “preserving”) and “crowdsourcing” and (“guarantees“ or “tasks” or “worker” or “surveys”). Similarly, for machine learning we use, “privacy” and “preservation” and “machine learning” and (“surveys” or “images” or “personal information”). The timeline of the literature spans from 2005 to 2019, and we capped our literature search to 30 search pages on average per keyword logic combination.

2.1 Privacy Preservation in Crowdsourcing

In this section, we first understand the effects of information extraction from the task content by the crowd. Based on the approaches used for privacy preservation, we organize the remaining parts of this section based on the obfuscation, segmentation, and task assignment methods.

2.1.1 Concerns of information extraction from task content

In crowdsourcing tasks, there is the possibility of the crowd workers being able to extract additional information from the task content which may not be necessary or relevant to complete the assigned task. If the task content is likely to contain some private information, the possibility of information extraction by the crowd concern is a security concern. Lasecki et al. [25] studied the effects on the task when the crowd workers extract information from the task content and manipulate their judgments by colluding task information amongst themselves.



(a) Scenario of workers extracting information from the task content [25]

(b) Scenario of workers manipulating their judgments for the task after colluding task information [25]

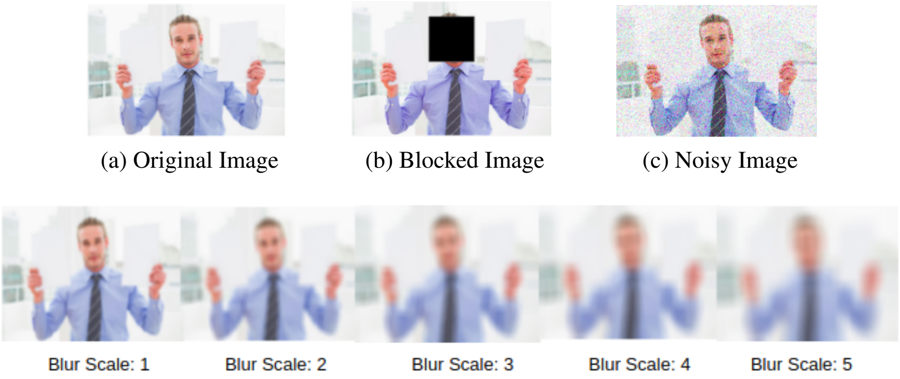
Figure 2.1: Concerns of information extraction from the task content in crowdsourcing tasks

Figure 2.1(a) represents a scenario of information extraction from the task content. In this image, the red worker has extracted information from the task’s image content, thus getting access to private information like credit card number, the name of the credit card holder from the visual cue. Thus, we define “information extraction” as an event when a crowd worker accidentally or intentionally extracts private information from the image used in the task. Figure 2.1(b) represents the scenario when workers collude information about the task and manipulate their judgments to alter or deviate from the expected output for the task. In this study, we focus on methods to limit the amount of private information visible to the crowd in tasks.

We understand the crowdsourcing works affected by information manipulation attacks are due to the susceptibility of information extraction from the task content [25]. We also observe that iterative crowdsourcing workflows are more prone to information extraction and manipulation attacks in each iterative update of the task. Manual verification of the task content for the amount of private information likely to be extracted can be a solution to limit or prevent extraction of private information in tasks. However, this solution becomes tedious and cumbersome for large image datasets. Our proposed hybrid approach removes manual verification for images by automating the detection of private visual cues in the image.

2.1.2 Task Content Obfuscation

Data Obfuscation is the process of de-identifying private visual cues through image processing methods like image blur. In this section, we discuss task content obfuscation through blurring [23, 22] and adding perturbations to distort images [41, 42]. Figure 2.2 gives an overview of the common obfuscation methods for de-identifying private visual cues in image task content.


(a) Original Image

(b) Blocked Image

(c) Noisy Image

(d) Image with progressive blurs

Figure 2.2: Summary of methods for de-identifying private visual cues in image task content.

Varshney et al. [41] were to our best knowledge, the first to propose the use of data perturbation methods like adding distortions in the image through visual noise for obfuscating private visual cues in the image. Figure 2.2(c) is an example of obfuscating private visual cues in the image by adding visual noise. [42] proposed an error-correcting codes based approach to determine the amount of visual noise added to obfuscate private visual cues in the image. Both [41, 42] study the trade-off between the privacy preserved by adding visual noise and the usefulness of the obfuscated image for the crowd to complete the crowdsourcing task. However, [41, 42] provide theoretical evaluations for the privacy preserved by adding visual noise to obfuscate private visual cues and cannot be considered as a representative for evaluations done on real crowdsourcing tasks.

Lasecki et al. study the impact of progressive blurring of private visual cues (shown in Figure 2.2(d)) impacts the accuracy of responses from the crowd while annotating behavioral videos [23]. The authors performed experiments by varying the level of blurring applied to the behavioral videos to find the optimal amount of blur which preserves the privacy of private visual cues and makes the obfuscated video frame useful for the crowd to annotate the video frame. From [42, 23] we find that the amount of privacy preserved in the task content by obfuscating private visual cues is inversely proportional to the usefulness of the task content for the crowd to complete tasks.

There are works which propose obfuscating private information in non-image task content like medical records [6]. If the task content has a predefined structure of organizing information, we can leverage the content structure to infer the location of private information. For medical records, [6] proposed obfuscating private information using a template. The template is made based on the fields containing private information in the medical record. We believe this obfuscation method can work on documents containing private information like bank statements, education records, and certificates.

2.1.3 Task Content Segmentation

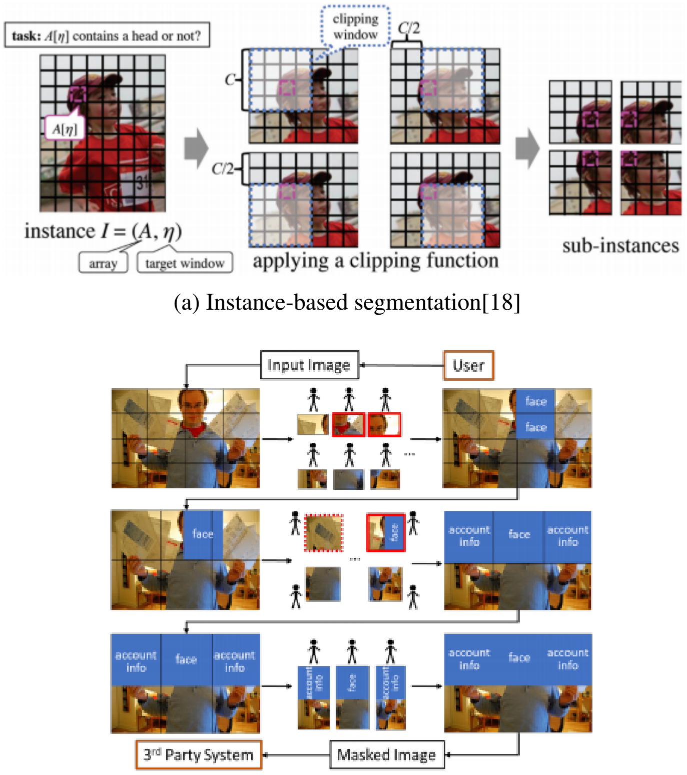
Through task content obfuscation, we can obfuscate some private information in the task content. However, it is still possible for the crowd to find private visual cues in the task content. To reduce the amount of private visual cues in the image, segmenting the image into smaller parts has been identified as a feasible approach in [18, 20].

Little et al. proposed one of the earliest task content segmentation approach for medical record annotating tasks [27]. There are two stages in their proposed task content segmentation approach. The first stage requires the crowd to annotate the fields of an empty medical record likely to contain personal information. In the second stage, the aggregated responses from the previous step are used to segment the medical record to minimize the number of private information visible to the crowd in each segment.

For images, task segmentation can be done in two approaches namely to: maximize privacy by limiting the segment size to be around a specific region on the image [18] as shown in Figure 2.3(a) and to progressively segment the image at different zoom levels to get fine-grained detection of private information in the image [20] as shown in Figure 2.3(b).

Kajino et al. [18] proposed an approach of creating non-overlapping image segments for regions of the image containing private visual cues. The segmentation approach uses a clipping window algorithm limiting the number of private information visible to the crowd in every segment. The evaluation is through a quantitative study on the crowd responses for the presence of private visual cues with and without the proposed segmentation approach.

Kaur et al. proposed an iterative crowdsourcing workflow for detecting private visual cues in images through segmentation [20]. Their proposed approach creates image segments based on the available budget and the desired amount of privacy to be preserved. The segment sizes progressively increase from small to large, with the crowdsourcing tasks created for detecting private visual cues for each increment of segment sizes. This work shows that the cost to preserve privacy is directly proportional



(a) Instance-based segmentation[18]

(b) Pyramid Workflow with progressive scaling of segment sizes[20]

Figure 2.3: Comparison of task content segmentation approaches

to the available budget for using the crowd for detecting private visual cues through the proposed pyramid workflow. Hence, a higher budget is required to maximize the privacy preserved in the image with a constant cost per unit task.

2.1.4 Task Assignment

Varshney et al. demonstrated privacy-preserving crowdsourcing task resilient to information extraction and manipulation attacks through workers colluding task information, shown in Figure 2.4 [41].

Celis et al. proposed a privacy-preserving task assignment approach assuming that the knowledge of the workers who collude task information is known [4]. The proposed task assignment approach segments the task content into task components. The assignment strategy ensures that two task components belonging to the same task content do not go to workers known to collude task information. The assignment strategy uses a function to estimate the amount of private information lost through workers colluding task information. The loss function provides guarantees on the amount of private information lost to colluding information.

In a recent work [39] proposed a three-step task assignment method which assumes that, instead of knowing the workers likely to collude information, it is possible to compute a pair-wise probability of the workers recruited to respond a particular

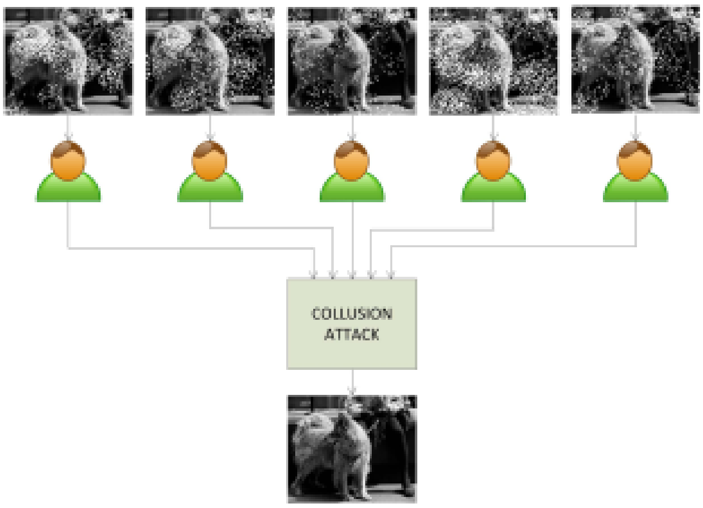


Figure 2.4: Information extraction from obfuscated image content through colluding workers[41].

crowdsourcing task. Workers who share the minimum pair-wise probability of colluding information would be assigned the task. In comparison to [4], the task assignment method proposed by [39] assumes a more realistic worker behavior as the computed pair-wise probability for recruited workers to collude task information is better than the general assumption of the workers who are likely to collude information is known.

However, if we were to look closely at the literature discussed in this section, we notice that [41, 4, 39] evaluate the task assignment methods on paper, leaving questions on implementation feasibility of these methods. The lack of uniform evaluation criteria and metrics for privacy preservation approaches in crowdsourcing makes comparing and evaluating claims on the amount of privacy preserved. The "loss function" from [4] which is defined based on the number of private information found in the content can be used as a standard evaluation metric. However, "loss function" as a metric requires knowledge of the private information present in the content indicating the need for a benchmark annotated dataset.

2.1.5 Privacy Preserving Crowd Applications and Systems

So far, we discussed privacy preservation methods applied to a crowdsourcing workflow. In this section, we discuss how real crowdsourcing applications preserve privacy in different task content like images, videos, and text.

Legion:AR [24] uses the crowd to annotate activities depicted in images using the crowd. Privacy preservation on the image by obfuscating of private visual cues like the people by covering the silhouettes of the people with a single color in the image. Obfuscating the people in the image prevents the crowd from extracting private information about the people and provides sufficient information to annotate the activity depicted in the image.

Zensors [22] is a crowdsourcing application used for real-time environmental sensing. This application uses images as the task content for sensing tasks, and the task requester can select the regions on the image and the obfuscation method for preserving privacy. Zensors++ [12] uses a state of the art machine learning model to detect and obfuscate faces of people in the image. Zensor++ uses low-resolution images, and videos without audio as additional measures for privacy preservation. We observe that [12] is an example of combining machine learning models used in crowdsourcing workflows for privacy preservation on the task content.

WearMail [40] is an email retrieval application with a conversational agent, chatbot for user interactions. Since email content private information like contact information, meeting information, confidential information related to an individual or work, it is necessary to preserve privacy on the email content. WearMail obfuscates sender names and their email addresses automatically, and private information appearing on the email's subject or body is obfuscated using a collection of protected words.

2.1.6 Summary

In this section on privacy preservation methods in crowdsourcing, we were able to understand the concern of information extraction attacks on the task content by the crowd [25]. However, we can limit the number of private visual cues in the image visible to the crowd for extracting information using data obfuscation methods like blurring the image task content [23] or by distorting the image using visual noise [42]. Task content obfuscation methods study the trade-off between the privacy preserved in the task content and the usefulness of the obfuscated image for the crowd to complete tasks.

Task content obfuscation approaches are still susceptible to workers extracting information by colluding task information. Thus, privacy-preserving task assignment methods [4, 39] can reduce the amount of private information extracted by workers colluding task information. However, the current state of the art privacy-preserving task assignment methods are theoretical and needs to be evaluated in real-crowdsourcing environments.

Task content segmentation methods like [18, 20] shows promising results for preserving privacy in task content, especially for images. While being effective methods for preserving privacy in image task content, these methods become expensive at scale. However, we believe combining task content obfuscation and segmentation methods can be an effective combination to preserve privacy in image task content.

Crowdsourcing applications and systems like [24, 22, 12] preserve privacy on the images used in for image analysis tasks using obfuscation methods like blurring and blocking private visual cues like the people in the image. To summarize, state of the art crowdsourcing workflows are either limited to the number of private information preserved in the task content or increases requester effort and expenses for crowdsourcing private information detecting tasks. In Table 2.1, we summarize the literature on privacy preservation methods used in crowdsourcing based on the type of content used in tasks and the method to preserve privacy on the content.

Content	Paper	Privacy Preservation Approach	Limitation
Image	[41, 42]	Uses data perturbation through visual noise to obfuscate private information visible in the task	Theoretical, Impact of visual noise on task accuracy needs to be measured
	[18]	Image is segmented into sub-regions, and the crowd annotates the private information in the image segment	Expensive at scale since the number of segments used in tasks increases
	[20]	Image is segmented in a pyramid workflow and the crowd iteratively annotates segments which increase in size	Number of segments created is driven by budget. The cost to preserve privacy is high, more small-sized segments are used
	[24]	Obfuscates the silhouettes of people in the image by covering it with a single color to preserve privacy	Other visual cues in the image considered as private information, is not obfuscated
	[22]	User selected region of the image and obfuscation method for the image allows users to create privacy-aware crowdsourcing tasks	Manually selection image regions for crowdsourcing becomes tedious and time-consuming for a large of collection of images, not scalable
Video, Image	[12]	Automates privacy preservation by obfuscating people‘s faces using face detection model and collects videos at low resolution with no audio. Introduces a hybrid privacy preservation scheme	Obfuscating faces is insufficient as other visual cues contain private information
Video	[23]	Studies the trade-off between privacy preserved and accuracy of the crowd for annotation tasks	Formal, deterministic method to compute the privacy vs accuracy trade-off is required
Image, Text	[6]	Uses a template to obfuscate private information in structured documents such as medical records	Effectiveness of the obfuscation depends on the structure of the task content matches to the template
	[4, 39]	Assumes that knowledge about workers who collude information is known and colluding workers are not assigned segments from the same content	Limited by implementation challenges, as it is not possible to get the knowledge of workers likely to collude information
	[27]	Based on the annotation of an empty medical record, the actual medical record is segmented and assigned for annotation task	Expensive at scale
Text	[40]	Automated obfuscation of private information in emails like email address, sender name before using the crowd for information finding tasks	Dictionary of private information needs to be prepared for specific use cases

Table 2.1: Summary of the literature on privacy preservation in crowdsourcing

2.2 Privacy Preservation through Machine Learning

Privacy preservation through machine learning is the application of machine learning or deep learning models to detect private information in different media. In this section, we discuss the different machine learning approaches to preserve privacy in textual documents and images.

2.2.1 User Generated Content on Social Media

User-generated social media content is used to learn about user behavior and is used to provide a personalized user experience on social media platforms. In recent years machine learning models have been used to preserve privacy on user-generated content like tweets, using state of the art methods like differential privacy. Differential privacy is a notion where an observer is unable to identify or differentiate if a particular individual's private information in a database. k-Anonymity is another popular privacy-preservation method which can provide privacy guarantees on information retrieved from a database. The challenges of preserving privacy on user-generated content lies on the unstructured nature and multiple media of the content, for example, tweets posted on Twitter, blog posts written on Medium and pictures shared on Instagram.

Song et al. propose a privacy-preserving approach which is based on the notion of privacy defined as what an individual share, when, to whom and under which circumstances [38]. Their approach has three components: the first component is used to collect data from Twitter and a taxonomy is created using the keywords associated with private information. The second component uses linguistic and metadata of Tweets as features to predict the personal information shared in the content using the previous taxonomy. Their model uses pre-defined features to detect private information from the taxonomy along with latent features specific to each private information to improve the prediction of the model. The third component provides alerts and suggestions to users regarding the steps to be taken in the event of their private information getting exposed. The suggestions are based on crowdsourced guidelines which are collected from the crowd using Amazon Mechanical Turk for a cross-cultural perspective on privacy preservation.

Attriguard [16] is a privacy preservation approach addressing automated inference attacks on web and mobile applications. Automated inference attacks refer to the use of machine learning models to extract user's private information like age, gender, location, political views from publicly shared data. Atriguard is based on the notion that that taxonomy-based privacy preservation approaches like [38] preserve privacy on the media at the cost of the utility of the content. The authors proposed an adversarial model which is used to find the minimum amount of random noise to be added to the content to defend from inference attacks. Attriguard aims to reduce the utility loss in the content without compromising on privacy by adding random noise to reduce the amount of private information extracted by the attacker. This work is an example of current works in privacy preservation through machine learning that focus on simulating inference attacks and also methods to minimize the amount of private information extracted through random perturbation through adversarial models.

2.2.2 For Text

Privacy preservation approaches for text through machine learning studies how machine learning models can automate the detection and preservation of private information in textual documents like clinical texts, application forms, user-reviews on social platforms and customer data. [47] is one of the early works on privacy preservation approaches for text where a cryptographic method is used to preserve privacy on customer data. This is one of the early works studying the trade-off between the trade of privacy preserved and the accuracy of training a data miner. In this section, we discuss a few works which preserve privacy on textual content on social media, clinical texts, and text sanitization approaches.

Clinical text

Electronic health records, insurance forms, clinical texts, bank statements are examples of textual documents which are likely to contain private information that can directly identify an individual. The extraction of private information from the mentioned examples of textual documents is using the pre-defined structure of the document which is used to locate and extract specific private information from the document.

In the healthcare domain, it is mandatory by law, for healthcare data providers to remove all the private information about individuals before sharing the data with researchers. [10] proposes the training of conditional random filter or CRF-based models to detect private information in Chinese clinical text. CRF-based models work well for detecting private information in clinical texts since the clinical text contain sequential data in a well-defined document structure. In general, CRF-based models are used for natural language processing tasks like named entity recognition, parts of speech tagging, etc. The proposed CRF-based model is used to detect and obfuscate private information like the name, age, address, health issues, medication provided to the individual. To train the model, a dataset from Chinese clinical records was manually processed and annotated to cover diverse categories of private information and the authors report comparable or slightly higher performance to the existing models based on the English language. We observe from the literature that the CRF-based models similar to [10] are effective in detecting private information from dense, structured text like application forms.

Text sanitization

Text sanitization refers to the process of removing sensitive or private information from textual documents. Redaction is the process of allowing selective information on the text to be available while obfuscating or blacking out sensitive and private information in the document [1]. Current works focus on the detection and obfuscation of private information in documents and also ensure that the documents are not completely obfuscated [9, 36].

Chow et al. proposed a model to detect private information, referred to as inferences in documents inspired from the association rule mining which states that inferences are based on word co-occurrences [9]. Their proposed model captures diverse

[1] https://en.wikipedia.org/wiki/Sanitization_(classified_information)

inferences in documents using a rule-based approach to find co-occurrences of private information along with a list of keywords associated with private information. The model approximates the knowledge about the private information of an individual using the web to preserve the individual's privacy.

Sanchez et al. proposed an automatic text sanitization approach which reduces human effort while ensuring that the text is useful [36]. The proposed sanitization model by [36] is based on a semantic privacy model which can detect private information in the text using features like keywords, concepts or word embeddings. Previous work-related to [36] had theoretical guarantees on the privacy preserved and usefulness of the text, but [36] allows different configurations for the trade-off between the degree of the privacy protected through sanitization and the utility of the document.

2.2.3 "Smart" Environments

"Smart" Environments are device ecosystems which aggregate data from different sources like sensor nodes and uses machine learning models to process the aggregated data to provide custom, personalized context-aware services. For example in assisted living environments, there are sensors which record and tracks the activity and health condition of the user. Assisted living systems can also be configured to contact emergency services, registered caregiver and relatives for sending specific information about the medical condition, activity along with their identity of the user. [33] notes that it is important to preserve the privacy of information about individuals living in ambient assisted living environments to ensure that the personalized, context-aware recommendations for the individual are processed using privacy-aware machine learning models.

Pyschoula et al. proposed a privacy-aware Long Short Term Memory (LSTM) model to encode and anonymize the data aggregated by devices in ambient assisted environments [33]. This means that the users can select specific people to whom they wish to share the aggregated data completely and the others to whom they prefer to send anonymous data. The proposed model creates different data views on the aggregated data which is a combination of specific private information about the individual and related generalized data. The LSTM model's encoder-decoder architecture is used to guarantee privacy as the data view would be useful only if the correct decoder is used. The authors also trained the proposed model to learn privacy operations based on the GDPR and state that the proposed approach could be used for preserving privacy in textual documents like clinical text, doctor's notes, etc.

2.2.4 For Images

In recent years, there have been significant advancements in deep learning models used in computer vision. For object recognition, YOLO [34] was fast and accurate in drawing bounding boxes around objects in videos and images. However, if we need to detect private information based on pixel-level information, the current pre-trained models can offer detection of certain visual cues only for example faces and the people in an image. We note the `VISPR` dataset used in [32] is a good dataset to evaluate privacy preservation approaches on images through machine learning since it has a diverse collection of annotated private information in images. In this section, we discuss current machine learning-based approaches to preserve privacy in images.

Current works on privacy preservation on images through machine learning or deep learning models are limited to specific private information like faces [8, 5], license plates[11] and study the trade-off between the privacy preserved and the utility or usefulness of the image.

Orekondy et al. proposed the first approach to detect a diverse category of private information in images through visual cues using machine learning models [31]. Their proposed approach used an ensemble of complex models each of which could detect private visual cues such as objects and text in the image. In this work, the authors study the trade-off between the amount of obfuscation for preserving privacy in the image and the usefulness of the image. The authors trained their proposed ensemble model on a pixel-level annotated dataset on private information in images, i.e. `VISPR` dataset. The pixel-level annotations were done manually by expert annotators for the 22k images from the `VISPR` dataset [32]. Preparing datasets on private information is done manually by a group expert annotators and hand-annotating large datasets is time-consuming and a cumbersome process.

Recently there has been an increasing interest to use generative adversarial networks (GANs) in a wide range of applications. GANs have also been used to preserve privacy in images [7] and for visual recognition in camera videos [46] where privacy is preserved by applying transformations on the original image and video feed respectively. In [7], the authors focus on images of human faces to demonstrate the effectiveness of GANs in preserving private information like ethnicity, gender, the identity of an individual which could be inferred from the face. [46] proposes an adversarial learning approach which optimizes the trade-off between the utility of the video for visual recognition tasks and privacy preserved. Their proposed application of adversarial learning is novel as the adversarial model acts against all models that can potentially extract private information from the video feed. Lastly, we see the application of the widely accepted and pragmatic machine learning approach to preserve privacy, *differential privacy*[2] being extended to preserve private information from visual cues like license plates [11].

Data Obfuscation

Data obfuscation approaches for privacy preservation through machine learning uses image transformation or perturbation methods like adding blurs to minimize the exposure of private information in the image, similar to the crowdsourcing approaches discussed in Section 2.1.2. In machine learning, data perturbation for obfuscation is done using adversarial models, for example, [29] adds perturbations through semantic segmentation of the image. Works like [26] point to another branch of data obfuscation research which study the impact of privacy preservation through data obfuscation methods like blurring or cartooning on the usefulness of the image. Specifically, [26, 31] explores the effectiveness of data obfuscation methods on human viewers, referring to the amount of private information that could be extracted or the usefulness and interpretability of the image by humans.

[2] https://en.wikipedia.org/wiki/Differential_privacy

Human Faces

Within privacy preservation on images through machine learning, the works preserving privacy on human faces has a collection of diverse approaches ranging from using adversarial perturbations [8] to cartooning [5]. In this section, we provide an overview of the different approaches to preserve privacy or obfuscate human faces in images.

The effectiveness of models used for automatic facial expression recognition reduces when the input images are privacy preserved through obfuscation methods like blurring. To ensure the image is useful for the model and at the same time preserve the facial expression and individual's privacy, [5] proposed an adversarial learning approach to learn "identity-invariant" features of the image and using surrogate replacement approaches where the original face is replaced by a realistic cartooned representation while preserving the facial expression.

Wang et al. proposed privacy preservation where machine learning models like the Support Vector Machines (SVM) is used as an external service for annotating images. Their proposed approach [45] uses semantic encryption on the image which is sent to the classifier, SVM. The SVM is trained on annotating encrypted images which ensures that the model does not learn from any private visual cues likely to be present in the image. This approach demonstrated how privacy-aware machine learning models can be offered as service to automatically annotate images. While this approach is secure, the exchange of encryption keys for large collections of images is a challenge.

The privacy threats and methods to preserve an individual's privacy from data analytics and information profiling on facial images are investigated in [8]. Their proposed approach preserves single or multiple facial attributes of an individual that can be used to directly identify the individual through adversarial perturbations. The proposed privacy preservation approach is based on the concept of *k-anonymity*[3] which is a widely accepted privacy preservation approach. The privacy preservation on the facial attributes is done such that the visual appearance of the image is not affected.

2.2.5 Summary

From the study of the current methods to preserve privacy through machine learning, we notice that privacy preservation done on text is usually for a dense, structured document where it is fairly easier to predict the fields of the documents that are likely to contain private information. In images, privacy preservation is achieved by a type of data obfuscation like blurring or cartooning and the discussed works which study the trade-off between the privacy preserved and the utility of the image.

Irrespective of whether privacy preservation is done on text or images, we observe that an annotated dataset is necessary and we see different datasets used in the discussed literature. The different image datasets used for privacy preservation makes it difficult for us to evaluate different privacy preservation approaches and we also note that most of these datasets have a limited scope of private information covered like covering only human faces, except for the `VISPR` dataset [32]. The `VISPR` dataset has a diverse coverage of private information in images, however, creating such datasets is time-consuming and cumbersome since they are generally manually annotated.

[3] https://en.wikipedia.org/wiki/K-anonymity

In summary, we note that privacy preservation approaches through machine learning for text is limited in detecting private information in an unstructured text which could be a study in natural language processing. The limitations for privacy preservation approaches for images is the limited coverage of private information, the use of an annotated dataset which limits the private information that could be detected by the proposed privacy-aware model and finally a standard dataset for evaluating the performance of the proposed privacy preservation approaches.

2.3 Conclusion

In this chapter, make a systematic study on the state of the art approaches to preserve privacy through crowdsourcing and machine learning and also briefly discuss the notion of hybrid human-machine approaches, leading to the first contribution of this thesis, C1.

In the first part of this study, we studied the different privacy preservation approaches through crowdsourcing. We find that current works study the trade-off between the privacy preserved and the cost or amount of obfuscation done (utility). For preserving privacy in images we note the data obfuscation methods like blurring [23] and segmenting the image into smaller tasks [18, 20] are feasible and practical. In recent works Zensors++ [12], we see the first instance of how machine learning models can be used in crowdsourcing workflows.

In the second part of this study, we studied the privacy preservation approaches through machine learning. Machine learning models are used to preserve privacy in documents as well as images. The models that preserve privacy in text are trained to detect private information on textual documents like clinical texts [10] and obfuscate the detected private information [36]. For images, we observe a lot of research has been done to preserve privacy in human faces through deidentification methods like cartooning [5] and blocking [31]. In images especially, there are very few works which can detect diverse private information from visual cues [31] and we attribute this limited coverage of private information to the lack of annotated datasets available for training these models. The machine learning approaches, we discussed in this chapter study the trade-off between the privacy preserved and the utility of the image. The common limitations for the discussed approaches are the requirements for large quantities of annotated data, and the absence of a benchmark datasets to evaluate the privacy preserved through different approaches. We think the `VISPR` dataset [32] can be used as a benchmark dataset in future works to evaluate privacy preservation approaches.

We observe that creating datasets with annotations for private information is done manually and it is a cumbersome and time-consuming process since models require large quantities of annotated images. This leads to a *vicious loop between machine learning and crowdsourcing* where crowdsourcing can be used to annotate images efficiently but there is the concern of private information getting leaked. Machine learning models cannot be used to preserve privacy without training the models on an annotated dataset.

The Research Gap

Based on our study on the current state of the art approaches for privacy preservation through machine learning and crowdsourcing, we identify the following *research gap*:

Machine learning models used in privacy preservation approaches requires a dataset which is annotated for the different private information, the availability of such datasets is limited [32]. In images, there are different datasets and models available which can preserve privacy by detecting and obfuscating either the person or the face of the person. However, there can be other visual cues in the image that can directly or indirectly identify the individual in the image. Thus, machine learning approaches alone is insufficient to detect private information in images. Crowdsourcing is a common technique used for image analysis tasks, and humans can annotate all possible visual cues in the image that likely contains private information. However, crowdsourcing tasks where humans are asked to detect private information needs to be segmented to ensure that the crowd has limited visibility to the private information in the image [18, 20]. Thus, using the crowd to detect private information in images becomes expensive at scale. This is a *vicious loop* that could be broken using hybrid human-machine approaches where both human and machine intelligence can be used to preserve privacy.

Hybrid human-machine approaches combine the strengths of machine learning and crowdsourcing to build solutions for tasks that neither machine learning nor crowdsourcing can be used individually [43]. Recent works which proposed the use of hybrid human-machine approaches in different domains like [21, 30, 2] have reported better performance results over approaches that use either machine learning or crowdsourcing only. Initially, Kamar et al. proposed an approach to optimize the hiring process of crowd workers by simultaneously reasoning about the uncertainties of the model and guide the data collection process from the crowd [19].

In a recent work, [21] studies how machine learning models and crowdsourcing can be combined to screen items, in this case, literature reviews that satisfy a set of predicates. The hybrid approach proposed by [21] for multi predicate screening uses a Partially Observable Markov Decision Process to adapt to specific attributes or feature of each item and the machine learning model is used to test different filters for the predicates and decide if the crowdsourcing tasks to collect screening filters for that item needs to be stopped. Their proposed approach takes the outputs of machine learning classifiers as priors to the class probability for each item. The prior class probabilities are refined by an adaptive crowdsourcing algorithm which is optimized for reducing the cost of asking the crowd. We take inspiration from this work to design a hybrid human-machine approach to detect private information in images.

According to [21], the design of hybrid approaches can be classified into two classes based on how the machine learning models are used. The first class of hybrid approaches uses machine learning models first and based on the confidence of the classifiers create crowdsourcing tasks. The second class of hybrid approaches uses machine learning models to generate filtering conditions which are verified and refined by the crowd. However, we find that the hybrid algorithm in [21] is based on prior probabilities, which in the case of detecting private information is not available. Our proposed hybrid algorithm uses confidence scores from machine learning models to determine the amount of crowdsourcing to be used for detecting private information. Using confidence scores from machine learning models also gives a more deterministic reason on how privacy is preserved in the image segments.

Chapter 3

Approach

From our study on the current approaches to preserve privacy in image content used in crowdsourcing tasks, we note that privacy preservation approaches through machine learning requires an annotated dataset to detect private information through visual cues in the image. Privacy preservation approaches through crowdsourcing are adaptable to detect private information in different contexts but are expensive when scaled over a large collection of images. Hybrid human-machine systems leverage the strengths of both machine learning and crowdsourcing to achieve more than either could achieve separately [43]. For example, [21] uses a Hybrid human-machine approach for item screening in literature survey and their results motivated us to consider using a hybrid human-machine approach to detect private information based on their performance in solving problems in different domains.

In this chapter, we propose a hybrid human-machine approach to detect private information which uses machine-generated image descriptions to create image segments for image analysis crowdsourcing, answering RSQ2. In addition to the proposed, hybrid human-machine approach, we make a mapping of visual instances in an image that describes private information.

3.1 Proposed Workflow

We propose a hybrid human-machine approach for detecting private information in an image called the "context-based segmentation". We use machine learning models to determine the amount of private information likely to be disclosed in the crowdsourcing tasks. We reduce the amount of private information disclosed through the image by segmenting the image using machine-generated descriptions which describes the image context. Figure 3.1, gives a high-level overview of the different stages of processing in the proposed hybrid human-machine workflow to reduce the number of private information in an image.

The input to the workflow is an image for which the privacy needs to be preserved. The first step of processing is to generate machine descriptions for the image shown in Figure 3.1(1). While machine learning models cannot detect private information in the image out-of-the-box, they can still be used to describe the different visual cues like the scene, objects, and text present in the image which can be used to compute if the image is likely to contain private information. Thus, we generate machine descriptions for the

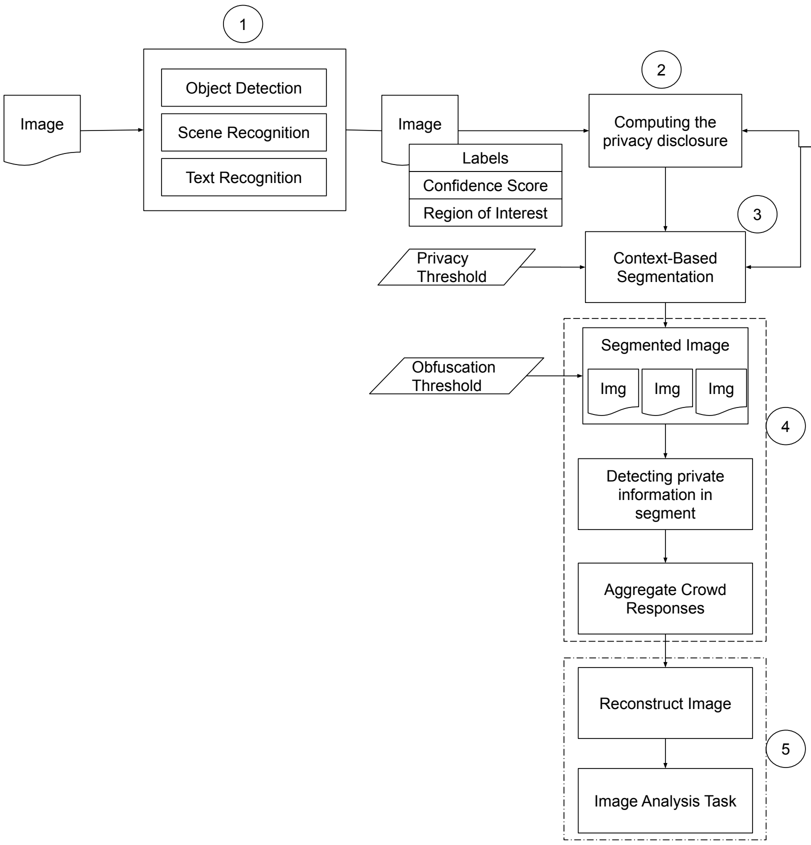


Figure 3.1: An overview of the proposed approach

image using an ensemble of machine learning models for object detection, scene, and text in the image. These models give a list of recognition machine generated instance labels with a value or score indicating the confidence of the model along with regions of interest on the image for which a particular instance label was generated by the models.

Based on the machine-generated descriptions, we compute the amount of private information present in the image based on a mapping of private information to the visual cues as shown in Figure 3.1(2). The computed amount of private information in the image is referred as "privacy disclosure" which is computed based on the labels and confidence scores outputs from each machine learning model for object detection, scene and text recognition. In this stage of processing, it is possible to obfuscate private information in the image related to an individual's like the gender, ethnicity, etc. that can be directly mapped to a machine-generated label.

Our proposed approach of context-based segmentation 3.1(3) uses the computed "privacy disclosure" for the image along with the "privacy threshold" to determine the size of the image segments used for the crowdsourcing to detect private information. The "privacy threshold" defines the acceptable levels of privacy disclosure computed for the image, and guides the proposed segmentation approach on the image with obfuscated individual private information like gender, ethnicity, etc.

The image segments are used in crowdsourcing tasks, where the crowd is asked to detect additional private information present in the image segment, represented in Figure 3.1(4). The fourth stage of processing takes as input an additional factor called the "obfuscation threshold" which defines the maximum amount of obfuscation done on the image. In short, the "obfuscation threshold" relates to the total number of black colored pixels in the image segment. The "obfuscation threshold" is used to study the trade-off between the privacy preserved and the usefulness of the image for the crowd [23, 31]. The "obfuscation threshold" impacts the number of segments sent to the crowd by defining the permissible amount of black colored pixels in the image segment, thus making the crowdsourcing tasks to detect private information in image segments cost-efficient.

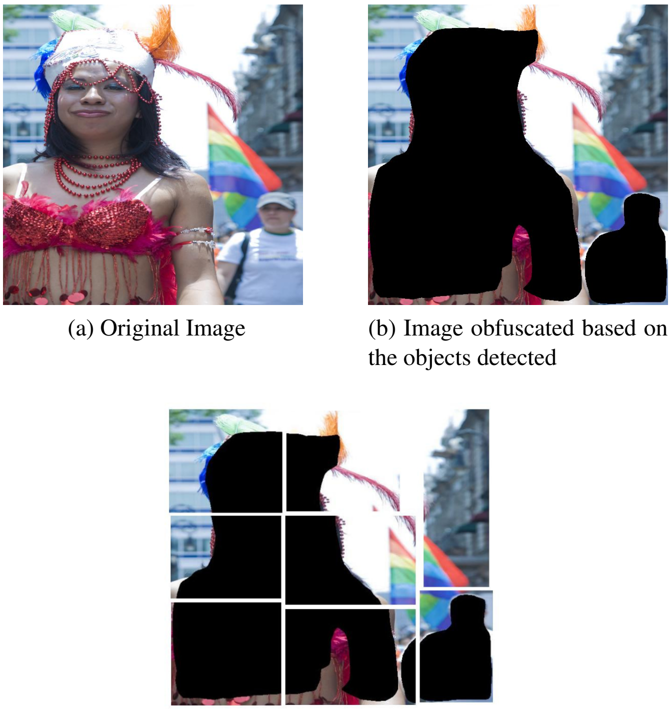
Based on the crowd annotations on the image segments, the image is reconstructed in the fifth stage of processing shown in Figure 3.1(5), where the reconstructed image is the obfuscated image where all private information in the image has been completely obfuscated. The reconstructed image is now used for the image analysis task like image annotation where the image is privacy-preserved and we study the trade-off between the privacy preserved and the usefulness of the image for the crowd [23].

3.2 The Notion of Context-based Segmentation

We illustrate an example of how our workflow creates privacy-preserved image segments for crowdsourcing tasks through Figure 3.2. To create a set of privacy-aware crowdsourcing task, the user uploads an image as input to our workflow shown in Figure 3.2(a) and cost specifications namely: the total budget to create crowdsourcing tasks for the given collection of images and the cost to be paid per task. In this example, assume the user sets $ 10 as the budget and $ 0.25 as the reward for the successful completion of a crowdsourcing task which will result in 40 crowdsourcing tasks.

Next, we use the pre-trained machine learning ensemble to identify private visual cues like the people in Figure 3.2(a) and obfuscate the detected object instances. The example image has private visual cues of a person, but if we were to segment the image without any obfuscation additional private information related to the people in the image such as their gender, sexual orientation along with socio-cultural factors like causes they support can be deduced by the crowd as shown in Figure 3.2(b) . This is the reason why we obfuscate private information before creating the image segments. At the same time, we get the scene attributes or scene descriptors to get the context for the image which in this case are `open area, touring, competing`.

Using the context for the image, we create image segments for crowdsourcing tasks. In this case, we assume that 5 crowd workers are needed to get accurate results from the crowd which means the maximum number of image segments that can be made for the example is 8 segments (40 crowdsourcing tasks for 5 different workers = 8 segments). This is where we see that the context of the image, as well as the detected region of obviously private visual cues, can be used to find the optimal segments for the crowdsourcing task. In this example, we note that the segments which contain people are smaller than the segment on the top-right corner in Figure 3.2(c) which does not contain any private visual cue. We also reject segments that are completely obfuscated, since all the pixels in that segment will have the same value. This is an automatic rejection condition which will save cost by limiting the number of crowdsourcing tasks created, so to the crowd we will send only 4 segments.



(a) Original Image

(b) Image obfuscated based on the objects detected

(c) Context-based segmentation based on the scene

Figure 3.2: Example of how context-based segmentation is done on an image

3.3 Mapping Private Information to Visual Cues in Images

To develop a method to detect private information in images, it is important to understand the different information that is considered private and are protected under privacy policies. It is also important to understand the visual cues and object instances in images that are likely to contain private information. From our study, we also find the annotated datasets can be replaced with a taxonomy of private information found in the image. In this section, we discuss our approach to build a mapping of private information to visual cues and object instances in the images inspired by the categorization of private information as shown in Figure 3.3.

The taxonomy is based on the different private information that can be found on multimedia content like travel dates, personal identifiers, contact details, location, financial details like credit card numbers, etc. We are inspired by the taxonomy structure of [35] and discuss our taxonomy of mapping visual cues in images to private information based on the GDPR in this section.

We started the process of mapping process by studying different policies on data protection and privacy. To ensure that we can cover a wide range of private information, our study spanned from data protection policies like the General Data Protection

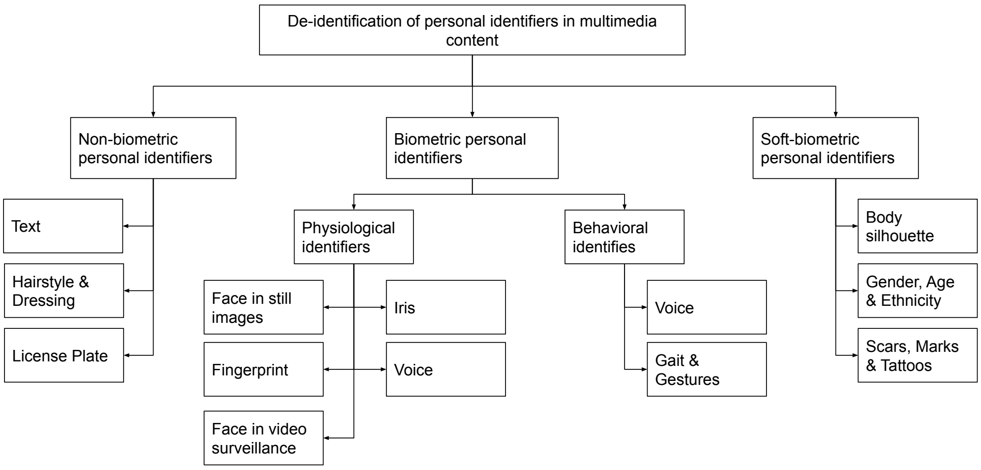

De-identification of personal identifiers in multimedia content
Non-biometric personal identifiers
Biometric personal identifiers
Soft-biometric personal identifiers
Text
Hairstyle & Dressing
License Plate
Physiological identifiers
Behavioral identifies
Face in still images
Iris
Fingerprint
Voice
Face in video surveillance
Voice
Gait & Gestures
Body silhouette
Gender, Age & Ethnicity
Scars, Marks & Tattoos

Figure 3.3: Taxonomy of personal information in multimedia content based on the 2016 EU privacy policy [35]

Regulation (GDPR)[1]. Among these policies, we find the GDPR to provide a clear categorization of private information, which forms the base for our mapping of private information. The GDPR define private information of an individual as "Personal Data" which include all information that can directly or indirectly identify a person. For instance, identification numbers are examples of information that directly identifies an individual and political opinions, religious opinions, financial status are information that can indirectly identify an individual. In our mapping shown in Figure 3.4, we categorize private information into seven categories which include directly identifiable private information like gender, ethnicity, biometric information of an individual that are mapped to visual cues like as well as private information like faces, photos, dresses as well as indirectly identifiable private information like the financial status, religion, religious opinions of an individual that are mapped to the visual cues like bank statements, credit card statements, religious objects, symbols and banners in the image.

The categorization of private information shown in Figure 3.4 is inspired by the different categories of data which the GDPR classifies as private[2]. This mapping of private information to visual cues and object instances in images is built following the information that is categorized as private in policies like the GDPR and forms the base for our proposed segmentation approach. The following are the categories of private information based on the mapping for which we give examples of visual cues that correspond to each category.

Identification Number contains private information that uniquely, directly identifies an individual. Typically license plates, citizen number(BSN), employee number, student registration number are examples of identification numbers that directly maps to the registration details about the vehicle, personal details like the name, address, education, etc. This information can be found on multiple object instances in the image like insurance policies, university transcripts, ID cards, passports. The category of iden-

[1] https://gdpr-info.eu/art-9-gdpr/
[2] https://gdpr-info.eu/art-4-gdpr/

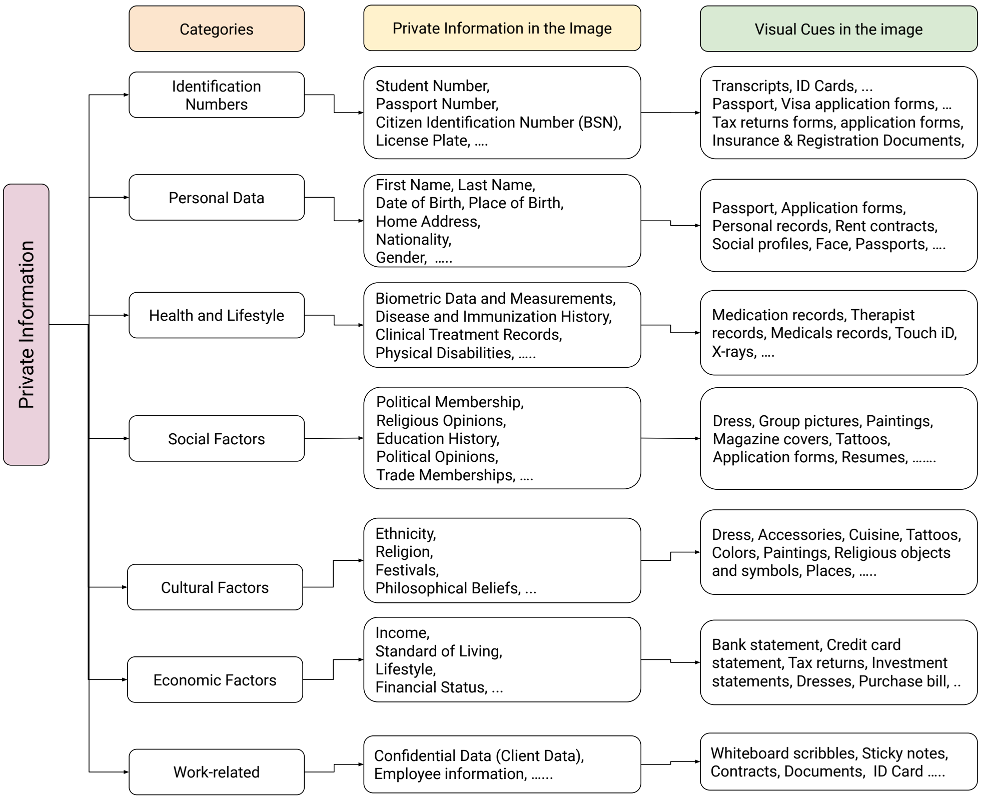

Categories
Private Information in the Image
Visual Cues in the image
Private Information
Identification Numbers
Student Number, Passport Number, Citizen Identification Number (BSN), License Plate, ...
Transcripts, ID Cards, ... Passport, Visa application forms, ... Tax returns forms, application forms, Insurance & Registration Documents,
Personal Data
First Name, Last Name, Date of Birth, Place of Birth, Home Address, Nationality, Gender, ...
Passport, Application forms, Personal records, Rent contracts, Social profiles, Face, Passports, ...
Health and Lifestyle
Biometric Data and Measurements, Disease and Immunization History, Clinical Treatment Records, Physical Disabilities, ...
Medication records, Therapist records, Medicals records, Touch iD, X-rays, ...
Social Factors
Political Membership, Religious Opinions, Education History, Political Opinions, Trade Memberships, ...
Dress, Group pictures, Paintings, Magazine covers, Tattoos, Application forms, Resumes, ...
Cultural Factors
Ethnicity, Religion, Festivals, Philosophical Beliefs, ...
Dress, Accessories, Cuisine, Tattoos, Colors, Paintings, Religious objects and symbols, Places, ...
Economic Factors
Income, Standard of Living, Lifestyle, Financial Status, ...
Bank statement, Credit card statement, Tax returns, Investment statements, Dresses, Purchase bill, ..
Work-related
Confidential Data (Client Data), Employee information, ...
Whiteboard scribbles, Sticky notes, Contracts, Documents, ID Card .....

Figure 3.4: An overview of the categorization of private information in our proposed mapping

tification numbers also includes Credit Card Number, Bank Account Number, Ticket (PNR) Number which can be used to extract private information about the spending, savings, and the travel itinerary planned for or by an individual which can be found in bank statements, credit card statements, and travel tickets.

Personal Data includes the private information which is traditionally processed as private information like the first name, last name, nationality, contact details, gender, family information, etc. This information can be extracted from object instances like passports, application forms, personal records, rent contracts. Personal data also includes factors like ethnicity, religion which can be inferred from instances of tattoos, symbols, and dresses and directly from personal data records.

Health and Lifestyle includes private information related to the health condition, healthcare treatments received and the personal lifestyle of an individual. Health conditions of an individual includes information like the mental health status, list of allergens, list of physical disabilities and immunization history which can be found from object instances like the personal health record, medical records, images of medical imagery X-ray and scans. Information about the healthcare information received by an individual can be found through object instances like the list of appointments with the therapist or doctor, medical or surgery records, medication record, disease history, medical prescriptions, and diagnostic reports. Lifestyle-related private information like the sexual orientation, mental health status can be found in object instance like the therapist notes, and treatment notes. Biometric data has also been included in this category since it is related to the human body and includes private information like the face geometry, fingerprints, unique physiological factors, etc. which are generally stored as images which are taken during different processes like during visa interviews and registration for citizenship.

Social Factors includes private information related to the qualifications, positions an individual holds in the society and the views and opinions shared by the individual. Information related to the qualifications of the individual corresponding to education, employments which can be found from professional social networks, resumes, application forms. The opinions and views of an individual are related to the political memberships, political views which can be inferred from object instances like magazines, banners, posters, flyers. The same object instances can be used to infer about the causes supported, and religious opinions of an individual. It is also possible to infer about an individual's social life through images of social gatherings and the dresses worn by the individual in those images. However, images of a gathering of people are usually under the freedom of the press act or are shared on public networks after getting individual consent.

Cultural Factors includes private information related to the ethnicity, religion, philosophical beliefs of an individual. Private information related to cultural factors is inferred through different object instances in an image. For example, the religion of a person could be inferred from object instances of known religious symbols, the ethnicity and philosophical of an individual could be deduced from festivals, dresses, paintings, etc.

Economic Factors include information that is related to the economic status of an individual like assets owned, the standard of living, financial status and income. This information can be extracted from tax returns filing documents, bank statements, credit card statements and can be inferred from object instances in an image like expensive objects used as interiors, well-known landmarks or distinctly identifiable landmarks, dresses, etc.

Work-related private information is usually confidential by nature can mostly be extracted from documents which contain information about the client, product, design or drawings of the product design, etc. An individual's type of work can also be extracted from the image of an ID Card which contains additional information that is related to the categories of Personal Data and Social Factors.

3.4 Context-Based Segmentation Algorithm

The Content-Based Segmentation is our proposed approach to create privacy-aware image segments for image analysis and is the contribution related to RSQ2.

Context-based segmentation is a hybrid human-machine approach where machine-generated image descriptors are obtained from machine learning models. These descriptors are logically compared with the mapping of private information and their corresponding visual cues to find the potential private information present in the image. The detected private information is obfuscated, and the obfuscated image is segmented and the segments are used in the crowdsourcing tasks to detect additional private information. The rationale behind naming the proposed approach “Context-based Segmentation” is the use of machine-generated descriptors to understand the contextual information in the image like the scene descriptions, and the list of objects and text present in the image. Since we obfuscate the private information before creating the crowdsourcing, we are likely to minimize the amount of private information visible to the crowd.

The Context-based segmentation approach is designed to preserve privacy for images, there are four additional inputs required to create privacy-preserving image segments for crowdsourcing tasks. The budget allocated to detect private information in the image (B), the cost per crowdsourcing tasks (C) and the number of workers ($N_{workers}$) required for reaching consensus are the initial inputs since the available budget determines the number of crowdsourcing tasks created. The cost per task is the minimum amount of money to be paid to the workers which depend on the type of task the crowd is required to complete and the pricing baselines suggested by crowdsourcing marketplaces. In addition to the user-specified inputs, the initialization step also involves loading the mapping of machine-generated instance labels for the image to the private information in the image denoted as P_{map}.

3.4.1 Generating Machine Descriptions of the Image

The first step in the context-based segmentation approach is to generate machine descriptions or instance labels of the image. The image descriptions or instance labels by themselves are not effective to detect private information present in the image but can be used to understand the context of the image. The context of the image can be defined as the image description which can be generated based on the scene, and the objects and text detected in the image. Thus, we use an ensemble of machine learning models to generate the descriptions and instance labels for the image, as shown in Algorithm 1.

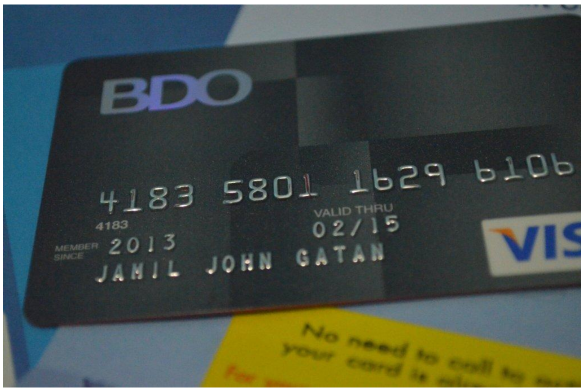
```
$[L_{object}, S_{object}, BB_{object}] \leftarrow Object\_Detection(\text{Image})$
$[L_{scene}, S_{scene}] \leftarrow Scene\_Recognition(\text{Image})$
$[L_{text}, S_{text}, BB_{text}] \leftarrow Text\_Recognition(\text{Image})$
```

Algorithm 1: Generating Machine Descriptions using an ensemble of machine learning models

Object Detection model is used to detect the objects or items along with their location on the image. The object detection model take the image I as input and outputs the a list of labels of the objects detected ($[L_{object}]$), along with the confidence score of the model ($[S_{object}]$) and the location of the detected objects in the image is specified as a region of interest through a list of bounding boxes ($[BB_{object}]$).

Scene Recognition model is used to get the scene descriptions for the image which are useful to understand whether the image is likely to contain some private information. Landscapes set in outdoors can most likely contain less private information than a scene of an office which has a high likelihood of containing private confidential or personal information about an individual or conveys the political or religious opinions of the individual. For an input image, I the scene recognition model returns a list of labels describing the scene ($[L_{scene}]$) and the confidence scores associated with each label ($[S_{scene}]$).

Text Recognition is a combination text detection and recognition models to identify the textual information in the text. Images with text can directly relate to private information belonging to the categories of personal data, work-related, health and lifestyle, economic and social factors. Given an image, the scene text detection and recognition models returns a list of tokens of the detected text ($[L_{text}]$), along with the confidence score of the model ($[S_{text}]$) and the location of the detected text in the image is specified as a region of interest through a list of bounding boxes ($[BB_{text}]$).



Object Detection: Credit Card, Table

Scene Recognition: Indoor room

Text Recognition: 'bdo', '4183', '5801', '1629', '6106', '4183', '2013', 'valid', 'jthru', '02115', 'member', 'since', '2013', 'jamil', 'john', 'gatan'

Figure 3.5: Example of machine generated labels for an image using object detection, scene and text recognition models

3.4.2 Computing the privacy disclosure for the image

From the machine-generated descriptors, we compute the privacy disclosure for the image based on the machine-generated instance labels from the object detection, scene and text recognition models using the mapping of private information to visual cues in the image P_{map}. The privacy disclosure for the object detection model (R_{Object}) is formulated as shown in Equation 3.1 which in can be defined as the ratio of the sum of the confidence scores of the object detection model to detect object instances ($[S_{object}]$) that relate to private information ($P_{map}(object)$) to the total number of object instances in the image which have a mapping to private information($|L_{object} \cap P_{map}(object)|$). The privacy disclosure for the scene recognition model shown in Equation 3.2 and scene OCR models are shown in Equation 3.3 is computed similar to the object detection model, but the scene recognition model uses the mapping of private information to scene labels ($P_{map}(scene)$) and the scene OCR models uses the mapping of private information to the detected tokens of text in the image ($P_{map}(text)$) respectively.

$$R_{object} \leftarrow \frac{\sum^{x \in L_{object} \cap P_{map}(object)} .s_x}{|L_{object} \cap P_{map}(object)|}, s_x \in S_{object} \tag{3.1}$$

$$R_{Scene} \leftarrow \frac{\sum^{x \in L_{scene} \cap P_{map}(scene)} .s_x}{|L_{scene} \cap P_{map}(scene)|}, s_x \in S_{scene} \tag{3.2}$$

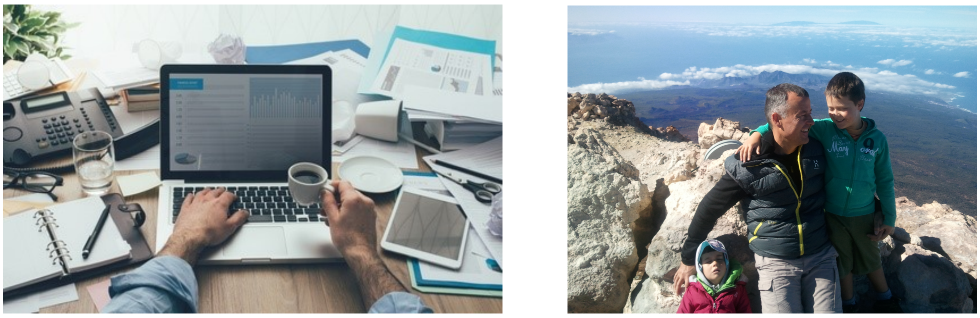
$$R_{Text} \leftarrow \frac{\sum^{x \in L_{text} \cap P_{map}(text)} .s_x}{|L_{text} \cap P_{map}(text)|}, s_x \in S_{text} \tag{3.3}$$

Obfuscating the image based on P_{map}

Using the mapping of private information to visual cues in the image (P_{map}), it is possible obfuscated the image based on the location of the region of the image containing the visual cues namely the object instances and the text specified by bounding boxes BB_{object}, BB_{text}. This results in the obfuscated image (I_{obfs}) where the private information that match to the mappings $P_{map}(object), P_{map}(text)$ are obfuscated based on machine-generated image descriptions L_{object}, L_{text}. This is step ensures that no obvious private information is visible to the crowd in the private information detection tasks that use the segmented image.

3.4.3 Context-based Segmentation Algorithm

The context-based segmentation is based on the notion of detecting private information effectively using the scene descriptions generated for an image. Consider the images shown in Figure 3.6, it is possible to deduce more private information from the indoor scene shown in Figure 3.6(a) than the outdoor scene shown in Figure 3.6(b). From Figure 3.6(a) it is possible to deduce private information related to the identity, nature of work and employment of an individual and sensitive or confidential information from visual cues like documents, emails or computer screens. Outdoor scenes like Figure 3.6(b) generally contain private information about an individual related to their gender, ethnicity, along with cultural and political opinions or license plates which can be mostly obfuscated using P_{map}. Thus, the number of segments made for Figure 3.6(a) should be more than Figure 3.6(b) to limit the amount of private information visible to the crowd per segment. The context to detect private information present in the image can be obtained based the machine generated image descriptors (L_{object}, L_{scene}, and L_{text}) and the mapping of private information to visual cues (P_{map}). The privacy disclosure for the image is computed individually for the object detection (R_{Object}), the scene recognition (R_{Scene}) and text recognition (R_{Text}) models to determine the total amount of private information that could be disclosed by the image. If the computed



(a) Example of an Indoor Scene

(b) Example of an Outdoor Scene

Figure 3.6: Difference in the amount of private information that can be extracted in indoor and outdoor images

total privacy disclosure of the image ($P_{computed}$) is represented as the ratio between the sum of privacy-disclosures of the object detection (R_{Object}) and text recognition (R_{Text}) models to the privacy disclosure of the scene recognition (R_{Scene}), the total privacy disclosure for the image ($P_{computed}$) will be always greater than 1. This increases the number of segments used to create crowdsourcing tasks to detect private information, resulting in proportional increase to the cost. Ideally, we would like the computed privacy disclosure for the image lie in the range $0 \leq P_{computed} \leq 1$, and so, we normalize the sum of R_{Object} and R_{Text} by the factor α as shown in Equation 3.4.

$$P_{computed} = \frac{\alpha.R_{Object} + (1-\alpha).R_{Text}}{R_{Scene}} \tag{3.4}$$

Based on the computed disclosure of private information by the image ($P_{computed}$), we can compute the size of the segments made by the context-based segmentation algorithm for creating privacy-aware crowdsourcing tasks. The inputs for the context-based segmentation algorithm are the budget for detecting private information per image (B), cost per crowdsourcing task (C), privacy threshold ($P_{threshold}$) which is an user-specified input value ranging from $[0,1]$ specifying the permissible disclosure of the detected private information in the image and the number of workers ($N_{workers}$) required for reaching consensus. The obfuscated image I_{obfs} from Section 3.4.2 is used for segmentation by first transforming the image into equal dimensions, such that it becomes a (mxm) square matrix. Thus, when the computed disclosure of private information by the image is less than the threshold, then the minimum number of possible segments is 4 where each segment has the dimension of $(\frac{m}{2}x\frac{m}{2})$.

```
Context-based segmentation($I_{obfs}$, $P_{computed}$, $P_{threshold}$, $B$, $C$, $N_{workers}$):
    if $P_{computed} \geq P_{threshold}$ then
        $numSegments \leftarrow \frac{B}{C}.\frac{P_{computed}}{N_{workers}}$
    else
        $numSegments \leftarrow 4$
    end
    $segmentDimension \leftarrow \sqrt{\frac{ImageArea}{numSegments}}$
    $[segments] \leftarrow segmentImage(I_{obfs}, segmentSize)$
    $createSegmentedTask([segments], B, C, O_{threshold})$
```

Algorithm 2: Context-Based Segmentation Algorithm

3.4.4 Creating tasks for the segmented image

For each image segment made through the proposed context-based segmentation, we ask the crowd to detect additional private information in the image segments. To determine which image segments would be sent to the crowd we measure the amount of obfuscation done on each segment ($O_{computed}$) and if it is less than the obfuscation threshold ($O_{threshold}$), we create the detection crowdsourcing tasks. Obfuscation Threshold ($O_{threshold}$) is an user-specified input value ranging from $[0,1]$ which specifies the amount of obfuscation that is permissible for image segments used for the crowdsourcing detection tasks. The method to create tasks for each segment of the image is described in Algorithm 3, where the image used is the obfuscated image I_{obfs} from Section 3.4.2. We compute the amount of obfuscation done on the segment $O_{computed}$ as the ratio of the total number of black colored pixels in the image segment to the total number of pixels in the image segment. We also ensure that not all pixels of the image segment is black, since creating crowdsourcing tasks to completely obfuscated or black colored image segments creates an additional cost of using the crowd.

```
createSegmentedTask([$segments$], $B$, $C$, $O_{threshold}$, $N_{workers}$):
    for s in segments do
        $O_{computed} \leftarrow \frac{s_{\text{black\_pixel}}}{s_{\text{total \_pixels\_in\_segment}}}$
        if $O_{computed} < O_{threshold}$ & $s_{all\_pixels}! = [0,0,0]$ then
            createCrowdsourcingTask(s)
    end
```

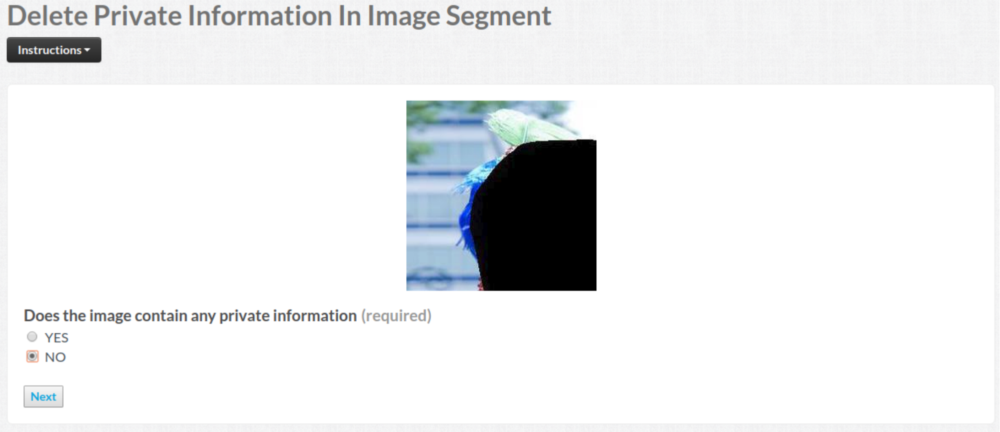
Algorithm 3: Creating Segmented Crowdsourcing Tasks

3.4.5 Design of detection crowdsourcing task

In the detection crowdsourcing task, we ask the crowd if the image segment shown contains additional private information which is not obfuscated. The task design is shown in Figure 3.7 where the crowd is asked to respond in a yes/no answering format for the presence of private information in the image segment, Figure 3.7(a). If the crowd answers “yes” for the presence of private information in the image segment then, we ask them to select the category of private information as shown in Figure 3.7(b). The categories of private information are based on the mapping of private information to visual cues in the image discussed in Section 3.3. About the task design choice, we opt for the “yes/no” answering format to understand the coverage of private information in the mapping of private information. We believe the responses from the crowd for the detection tasks give a preliminary estimation of the effectiveness of the context-based segmentation approach for privacy preservation.

3.5 Summary

In this chapter, we describe and elaborate on the different stages of our proposed context-based segmentation approach of creating privacy-aware crowdsourcing tasks to detect private information in images. The formal definitions and theory behind the thesis contributions C2 in which we develop a taxonomy to map different visual cues in the image to private information and C3 which corresponds to the definition of the context-based segmentation algorithm.

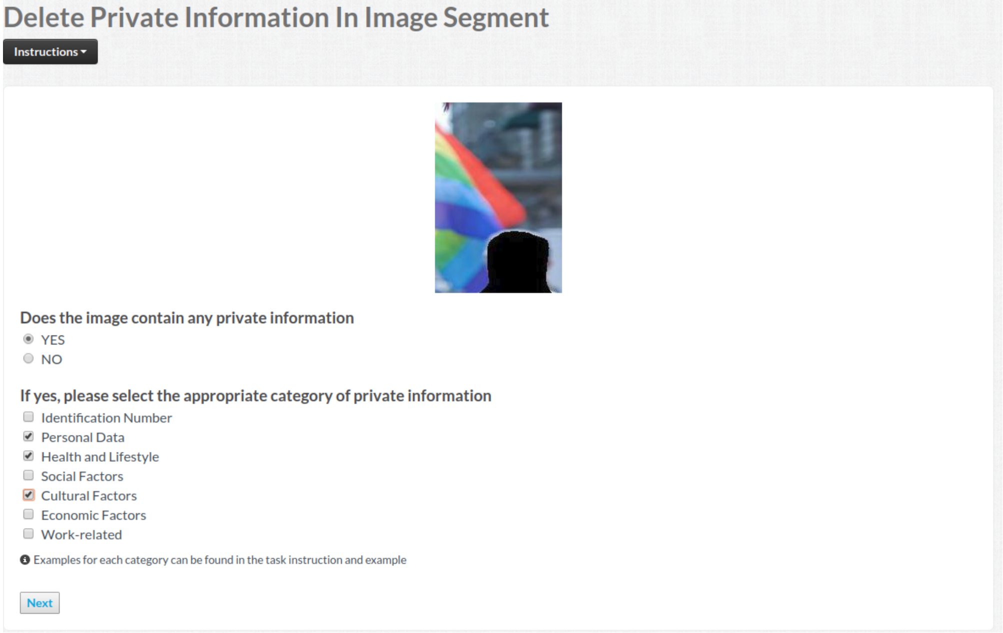


Delete Private Information In Image Segment

Does the image contain any private information (required)

○ YES
● NO

(a) Example where no additional private information is found



Delete Private Information In Image Segment

Does the image contain any private information

● YES
○ NO

If yes, please select the appropriate category of private information

- ☐ Identification Number
- ☑ Personal Data
- ☑ Health and Lifestyle
- ☐ Social Factors
- ☑ Cultural Factors
- ☐ Economic Factors
- ☐ Work-related

ℹ Examples for each category can be found in the task instruction and example

(b) Example where the segment contains private information

Figure 3.7: Design of the crowdsourcing task to detect private information in image segments

Chapter 4

Implementation

In this chapter, we discuss the choices of models made to make the ensemble of machine learning models used to generate machine descriptions for the image, along with the implementation of the mapping of private information to the visual cues in an image, the design for the the different crowdsourcing tasks and deployments of the proposed workflow.

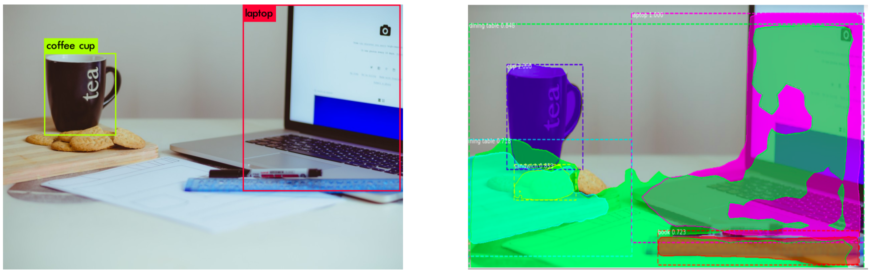
4.1 Implementing the Machine Learning Ensemble

Generating machine descriptions for the image is the first stage of our proposed workflow and we use an ensemble of different pre-trained, deep learning models to generate the image descriptions. In this section, we discuss the design choices made for selecting the models used for object detection, scene recognition, and scene text OCR.

4.1.1 Object Detection

Object Detection models are widely used in computer vision and there are different options to select a suitable object detection model for a specific use-case. From the pool of the available object detection approaches, we narrow-down to two popular models namely the “You Only Look Once” (YOLO) model [34] and the mask R-CNN model [1]. We were able to scope down the models based on the underlying design of the object detection network, in this case, the recurrent convolutional neural network (R-CNN) which perform very well for visual perspective tasks like semantic segmentation, instance segmentation, etc.

With Mask-RCNN [14] being made available in 2017, there have been improved models for object detection[1]. The particular implementation [1] is based on the implementation of the Mask R-CNN [14] and is a two-stage object detection process where in the first stage, the network scans the image to generate proposals of regions of interest in the image. Regions of interest in the image relate to the location of the object in the bounding box and are typically represented through a bounding box. The process of generating regions of interest of objects in the image through bounding box is the underlying framework used in most object detection approaches like Faster R-CNN, YOLO, etc. The second stage of the Mask R-CNN approach for object detection is to generate masks for the detected objects where a convolutional network takes the bounding box or region of interest as an input to generate the object mask.



(a) Bounding boxes with object labels from YOLO

(b) Bounding Boxes, Object Masks and Object Labels from Mask R-CNN

Figure 4.1: Comparison of the object detection models considered for implementation

For our implementation, we use the Mask R-CNN approach of object detection [1] which combines the use of instance segmentation and object detection. Image segmentation is defined as the task to identify the outlines for detected objects in the pixel-level [1]. In object detection, there are two popular segmentation types namely semantic segmentation which detects unique labels of objects in the image and instance segmentation which gives a count for each detected object class. As a requirement, we wanted to count the number of detected classes of object instances in the image and thus preferred the instance segmentation based object detection based [1] for our implementation. The choice between YOLO[34] and Mask R-CNN[14] was made in favor of the Mask R-CNN approach since it gives flexibility in obfuscating the detected objects either based on regions (bounding boxes) or object masks, illustrated in Figure 4.1. We also verify the reported performance of the pre-trained YOLO and Mask R-CNN models for object detection on the Microsoft COCO dataset [2], and finalize the Mask R-CNN based [1] over YOLO.

4.1.2 Scene Recognition

The choice of using the scene recognition model is to get scene descriptors for the image based on the scene descriptors the context-based segmentation algorithm determines the segment size for segmenting the image to be used in image analysis tasks. We chose [49], which is a pre-trained model to recognize the scene depicted by the image where the model returns a list of scene attributes that describes the image. From extensive testing on a collection of images, we found that the scene attributes for the image were able to give a more fine-grained description for the image which significantly improves the performance of the proposed context-based segmentation approach. The image attributes are stored as a list of adjectives which describe the different actions likely to be performed by the actors for instance reading, socializing, dining, congregating and scene descriptors man-made, open-area, closed-area, no-horizon. The scene detection model is also based on the recurrent convolutional

[1] https://engineering.matterport.com/splash-of-color-instance-segmentation-with-mask-r-cnn-and-tensorflow-7c761e238b46

[2] http://cocodataset.org/#home

neural network (R-CNN) architecture and is trained on a dataset of close to 1000 images [3].

4.1.3 Text Recognition

The choice of models to detect text in images was based on how well the model was able to detect all the textual information in the scene described in an image. The textual information in the scene includes but is not limited to the text of direction boards, advertisement hoardings, warning boards in outdoor scenes and text on handwritten notes, whiteboard, or printed text on the screens and papers in indoor scenes. We tested all the models on a collection of images of indoor and outdoor scenes with textual information and used it to benchmark the models for text recognition. Popular text recognition models that are based on Long Short-term Memory models are widely used for optical character recognition (OCR) applications like Tesseract. From our tests, we were able to observe that the LSTM based approach was able to preserve the order of text and worked very well for text on printed documents and the images of scanned documents but did not recognize the text in images of outdoor scenes and thus was not considered for the implementation.

We decided to combine two pre-trained models where one model is used to detect the text in the image irrespective of the scene by specifying the regions of interest (bounding boxes) and another model can recognize the textual information in the detected regions of interests on the image. For the task of detecting text in the image irrespective of the scene where we considered two state-of-the-art approaches namely FOTS[28] and EAST[50]. We decided to use the EAST model[4] for the implementation since there was a pre-trained model available which had a stable performance throughout the tests for model selection, FOTS which had better performance results on paper was not used since we were unable to find a stable implementation with pre-trained weights. At the time of model selection, we found a Mask R-CNN based text detection approach [15] which outperforms the model, currently used in the implementation but at the time of implementation we were unable to find pre-trained weights and the source code for the model. For recognizing the text from the detected regions of the text, we use a convolutional neural network model from [37]. We also considered license plates detection as a separate model and tested OpenALPR library [5] which works well to detect license plates by defining a region (bounding box) on the image.

4.2 Implementing the Taxonomy of Private Information in Images

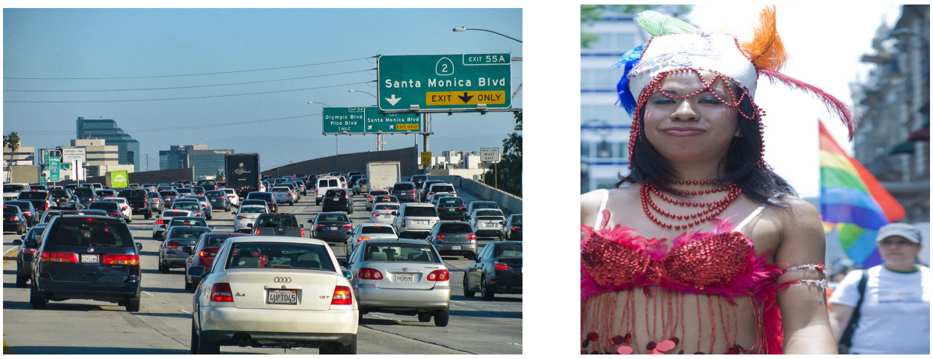
Before segmenting the image based on scene descriptors, we obfuscate some private information in the image using a taxonomy or mapping of the private information to the visual cues in the image. In this section, we describe the implementation of the taxonomy of private information in images which were theoretically discussed in Section 3.3 and this taxonomy corresponds to the thesis contribution, C2.

[3] https://places2.csail.mit.edu

[4] https://github.com/argman/EAST

[5] https://github.com/openalpr/openalpr

We build the taxonomy of private information in images using the categories derived from the GDPR and associate objects or visual cues that are likely to contain private information to the machine-generated labels for the detected object, scene, and text in the image. The object descriptors generated by the object detection model [1] are the annotated object classes of the Microsoft COCO dataset. For example, the private information related to an individual like the gender, ethnicity, physical disabilities can be inferred from the image of the person in the image, and thus are mapped to the detected object label "`person`". This is an example when the detected private information directly maps to the visual cue in the taxonomy, however, when we consider the example an ID Card which contains different private information about an individual is classified as a "`book`" and the photo of the individual is classified as a "`person`". An ID Card can be detected by applying a logical rule as, "if the person is detected within the book's bounding box area, then the visual cue can be an ID card". However, it is not possible to maximize the coverage of private information using a rule-based approach. Thus, we combine the detected object classes with the machine-generated labels from the scene and text recognition models. For the scene descriptors we map private information based on the "scene categories", "scene attributes" and "scene environment" generated from the implementation of [49]. The "scene category" and "scene environment" gives a broad understanding of the scene and the private information likely to be found in the image. However the additional image descriptions, "scene attributes" predicted by the scene recognition model of [49] helps us to get a better understanding of the different visual cues present in the image, which includes objects not detected by the object detection model. Thus, we use the "scene category" and "scene environment" to compute the privacy disclosure for a scene and use the "scene attributes" to optimize the number of segments produced by the context-based segmentation algorithm. Consider the images shown in Figure 4.2(b) it is possible to deduce private information like the sexual orientation and cultural opinions of an individual, but Figure 4.2(a) which is also an outdoor scene has considerably less deducible private information.



(a) an outdoor scene with limited private information

(b) an outdoor scene with people with more private information

Figure 4.2: Example of the different private information that can be deduced from outdoor scenes

For the detected text by the text recognition model, we create a mapping for (i) `name` by checking if it exists in a list of 613K first names or 162K last names obtained from the US Census Bureau website, (ii) `nationality` by checking if it exists in a list of 193 countries and nationality names, (iii) `location` if it exists in the list of 189K city names taken from the GeoNames geographical database. It is possible to map private information like license plates, email addresses with a regular expression based rule, however, the text are recognized as tokens and the generated tokens are not in the same order as the text in the image. However, when we implemented the mapping based on regular expressions we got inconsistent performance in text obfuscation on the image which is attributed to the machine learning models.

4.3 Summary

We develop the prototype of the context-based segmentation algorithm which works on a directory on images as input, with additional inputs like the total budget for using the crowd to detect private information in all the images in the input directory along with the cost per crowdsourcing task, number of workers and the threshold values for the amount of private information disclosed by the image and the amount of obfuscation permitted per image segment.

The prototype uses a pre-processed taxonomy of private information in images, which is stored as a dictionary and additional data sources like the list of person names, country names are stored as separate text files. The dictionary stores the following information: the category of private information, the visual cue which corresponds to the private information, along with the machine-generated scene and object labels. For some visual cues, like credit cards which have a standard structure, we bind a set of regular expressions to map the detected text in the image, in this case: the presence of 4x 4-digit numbers, 2x dates in MM/YY format. In the dictionary, each record corresponds to one visual cue which can be associated with one or more than one category of private information.

We use Figure Eight as the crowdsourcing marketplace to collect annotations on the additional private information present in the image segment. We automate the task creation for all the images using the same budget for all the images in the directory. Based on the aggregated responses from the crowd, we reconstruct the image and use it for image analysis crowdsourcing tasks. For our experiments, we used Amazon Mechanical Turk since we ran out of the free task limit in Figure Eight.

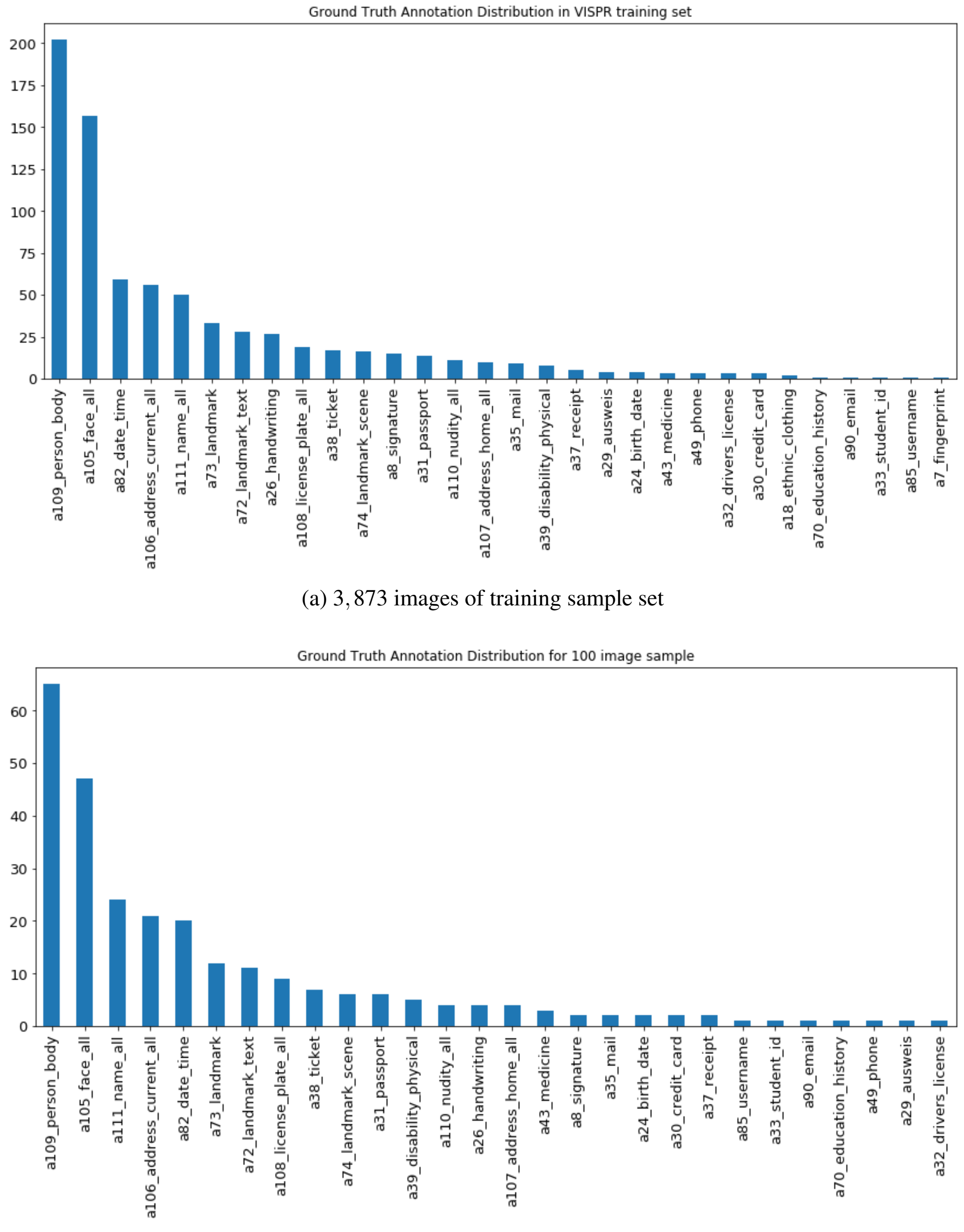
Chapter 5

Experiments and Results

In this chapter, we start by discussing the dataset used in our experiments, followed by the different parameters and experiments used to evaluate the proposed context-based segmentation approach for preserving privacy in images.

5.1 Dataset

In our literature study, we observed that the current state of the art methods for privacy preservation in both machine learning and crowdsourcing use different datasets for performance evaluation and validation. However, most of these datasets were for a particular private visual cue in the image like an individual's face or body [8, 5]. We wanted to evaluate the performance of our proposed privacy preservation approach through context-based segmentation with a dataset annotated for a heterogeneous mix of private visual cues in the image. The `Viz-Wiz` dataset by [3] which was used in [20] or the Stanford 40 Action dataset [48] used in [18] were not sufficiently annotated for private visual cues. We performed our experiments with the `VISPR` dataset containing over 8,000 images annotated over 24 different visual cues likely to contain private information [31]. This year, Gurari et al. released an updated version of the `Viz-Wiz` dataset [13] with proposed increased in the number of annotated private visual cues. However, we found that the `VISPR` dataset had more diverse categories for annotating private visual cues and hence used it in our experiments. For creating the dataset for our experiments, we performed an exploratory analysis over the train, test, and validation parts of the `VISPR` dataset [31]. The distribution of annotated labels observed a pattern similar to the frequency distribution shown in Figure 5.1(a) across the `VISPR` dataset. The result of the exploratory analysis indicated a strong skewness in the distribution of the ground truth between the `person_body` and `face_all` labels. We randomly sample 100 images from the 3,873 images in the training set. Even after performing the random sampling, we find the skewness in the distribution of ground truth labels shown in Figure 5.1(b) is similar to Figure 5.1(a). We bifurcate the `landmark` class label based on the visual cues present as `landmark_text` for images containing visual cues of street signs, addresses, advertisements, and `landmark_scene` for images of monuments, buildings and places of worship. We use this subset of 100 images for the experiments discussed in this section.



(a) 3,873 images of training sample set

(b) 100 images used in our experiments

Figure 5.1: Frequency distribution of private annotations on the images in the `VISPR` dataset

5.2 Experimental Setup

We perform three experiments with the common objective to measure the amount of private information detected in the image or image segments. The first experiment uses crowdsourcing-only approaches for privacy preservation, where we segment the image and ask the crowd to detect private information in the image segments. In this experiment, we can do only a qualitative study for the amount of private information detected in the image segment since there are no segment-level ground truth labels. The second experiment uses only machine learning models to detect private information, and we evaluate the amount of private information detected correctly by comparing with the ground truth annotations for the image through quantitative measures like precision and recall. In the third experiment, we evaluate the amount of privacy preserved in the image segments created using the context-based segmentation through a qualitative study on the responses from the crowd. In this experiment, we test the amount of privacy preserved by varying the threshold values for the computed privacy and computed obfuscation values.

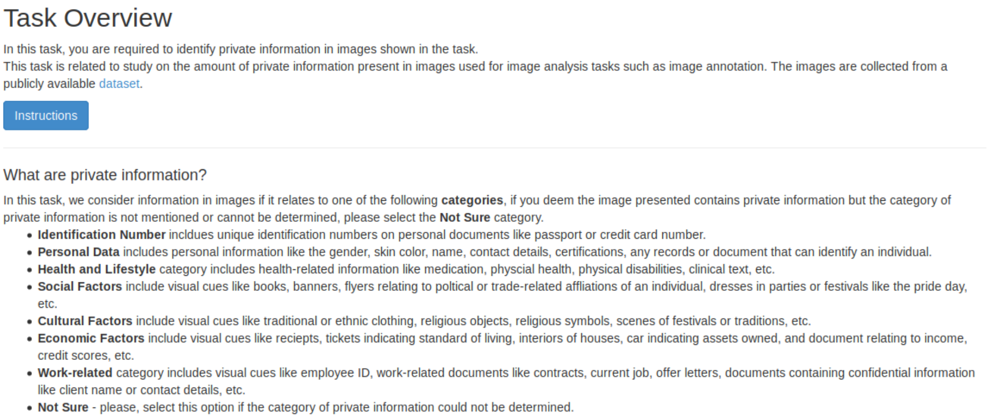
5.3 Experiment 1: Crowdsourcing Only

In this experiment, our objective is to study the amount of private information that can be preserved by segmenting the image for the image analysis task.

In this experiment, we study how much private information is detected by the crowd when the image is segmented. From the results reported by the state of the art approaches for segmenting the task content, we observe that the smaller the size of the image segment, the more privacy is preserved [18, 20]. In our experiment, we validate the hypothesis if smaller segment sizes are effective in reducing the amount of private information in the image segment. To test this hypothesis, we create three-sizes of non-overlapping image segments using a clipping algorithm similar to [18]. Thus, for one image we create 16 small-sized segments, 9 medium-sized segments, and 4 large-sized segments. This means one image has 29 image segments and in total there are 2900 image segments for our dataset of 100 images.

We created 290 crowdsourcing tasks from the 2900 image segments, such that there are 10 image segments per task. We randomly sample 10 images from the sample space of 2900 image segments, to reduce the number of image segments corresponding to the same image in one task. This random sampling also ensures the workers get to see different images in the task, making it less monotonous.

There are two UI components in the crowdsourcing task to detect private visual cues in the given image segment. The first UI component gives workers a list of all the categories of private information relevant to the task with examples to visual cues containing private information belonging to each category, shown in Figure 5.2(a). The categories of private information correspond to the categories of private visual cues in the mapping discussed in Section 3.2. We also designed the examples for the task using a pop-up window interface for a button click event, to make it easy for workers to work on the task. The second UI component is the actual crowdsourcing task which contains a container to display the image segment for which we need the crowd to detect the presence of any private visual cues, shown in Figure 5.2(b).



Task Overview

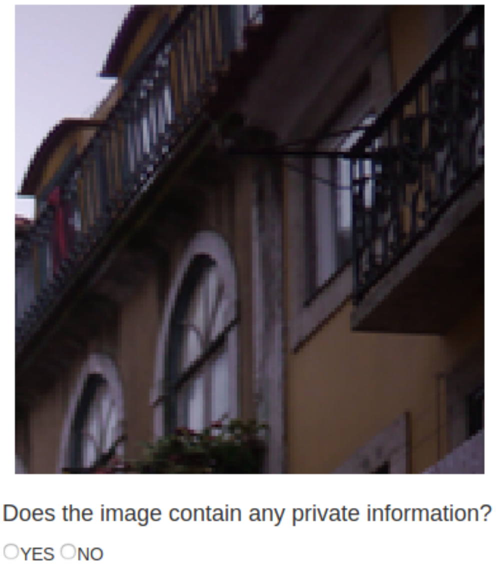
In this task, you are required to identify private information in images shown in the task.
This task is related to study on the amount of private information present in images used for image analysis tasks such as image annotation. The images are collected from a publicly available dataset.

What are private information?

In this task, we consider information in images if it relates to one of the following **categories**, if you deem the image presented contains private information but the category of private information is not mentioned or cannot be determined, please select the **Not Sure** category.

- **Identification Number** incldues unique identification numbers on personal documents like passport or credit card number.
- **Personal Data** includes personal information like the gender, skin color, name, contact details, certifications, any records or document that can identify an individual.
- **Health and Lifestyle** category includes health-related information like medication, physcial health, physical disabilities, clinical text, etc.
- **Social Factors** include visual cues like books, banners, flyers relating to poltical or trade-related affliations of an individual, dresses in parties or festivals like the pride day, etc.
- **Cultural Factors** include visual cues like traditional or ethnic clothing, religious objects, religious symbols, scenes of festivals or traditions, etc.
- **Economic Factors** include visual cues like reciepts, tickets indicating standard of living, interiors of houses, car indicating assets owned, and document relating to income, credit scores, etc.
- **Work-related** category includes visual cues like employee ID, work-related documents like contracts, current job, offer letters, documents containing confidential information like client name or contact details, etc.
- **Not Sure** - please, select this option if the category of private information could not be determined.

(a) Task instruction



Does the image contain any private information?

○YES ○NO

(b) Initialized task UI

Figure 5.2: Initialization of the Crowdsourcing only experiments on Amazon Mechanical Turk

We run the 290 crowdsourcing tasks for detecting and classifying private visual cues in the image segments on Amazon Mechanical Turk. We set $0.02 as the incentive for detecting private visual cues for one image segment. We aggregate the responses for the tasks from three workers using the majority voting aggregation method. The first question in the detection task asks the crowd to detect if the image segment contains private visual cues similar to the ones mentioned in the task instruction. Based on the response from the worker, if there are **NO** private visual cues in the segment, shown in Figure 5.3(a), the button to view the next task is enabled. If the worker indicates that **YES** private visual cues are present in the image segment, we shoe the interface for the workers to select the categories of private information to which the detected visual cues match as shown in Figure 5.3(b). Only after the workers indicate at least one category of private information will they be able to see the next image in the task.

The aggregated responses from the crowd are analyzed and we find that close to 55% of the image segments do not contain private visual cues, as shown in Figure 5.4(a). In terms of the 45% of the image segments which contained private information, we observe that `personal_data` category appears more frequently than the other categories of private information, as shown in Figure 5.4(b). The frequency distribution indicates there is a lot of visual cues belonging to the `personal_data` category like a mobile phone or laptop screens, ID cards, a person in the picture disclosing private information like the person's name, contact details, fingerprints, ethnicity, and cultural or social identity.

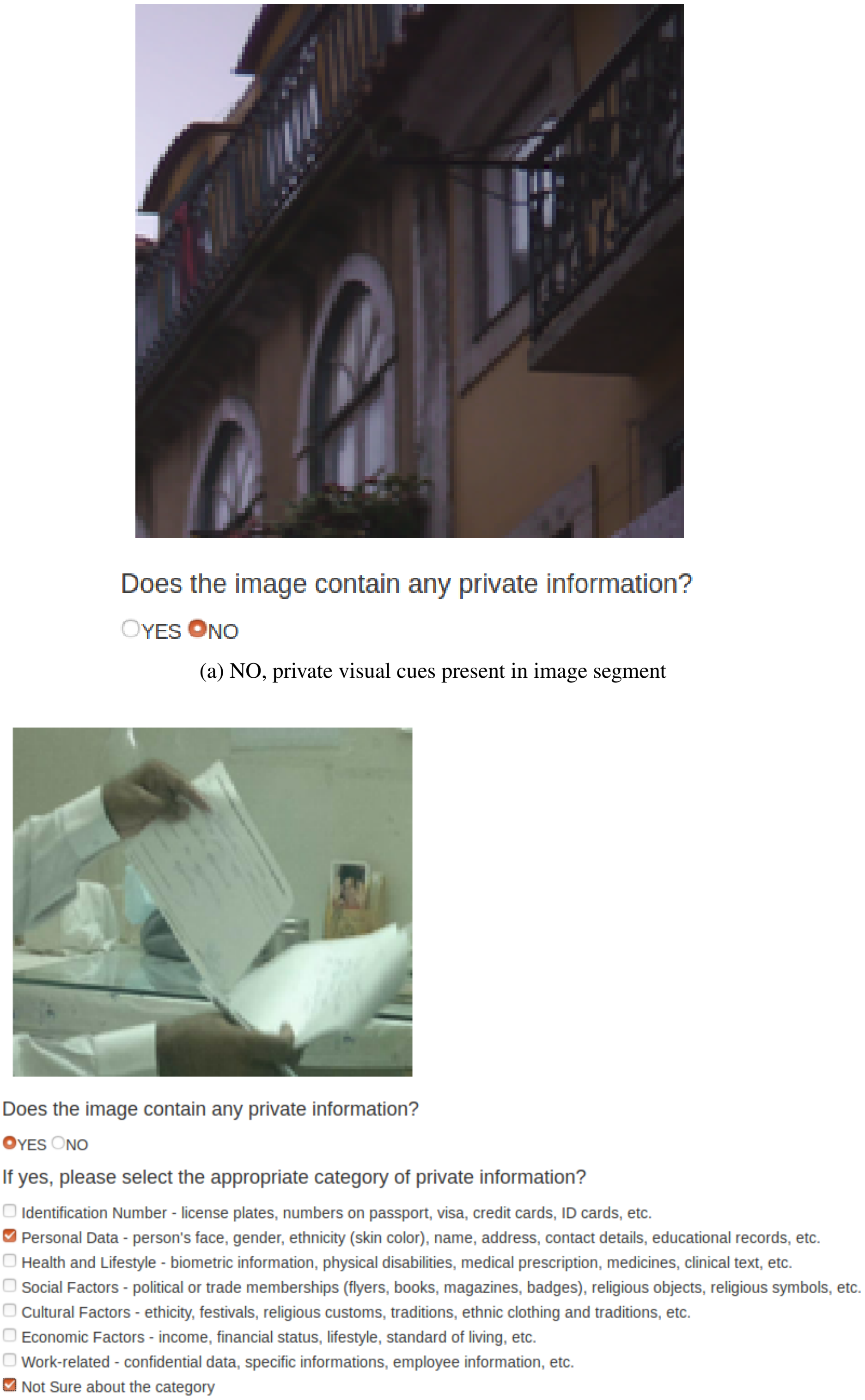
In some image segments, visual cues of the clothing worn by people in the image disclose private information related to `social_factors` like sexual orientation, causes supported or `cultural_factors` like ethnicity, religious beliefs of the individual. We note that there is a large overlap between `social_factors` and `cultural_factors` categories of private information in the responses as cultural events like weddings and social events like addressing a rally or a pride day party, were misclassified. We also noticed that workers are unclear when to use and not use the `not_sure` category and find classifying images containing segments of certain visual cues like tickets, and human finger difficult.

In this experiment, we are unable to use quantitative measures like precision and recall for evaluating the responses from the crowd because of the lack of segment-level ground truth annotations. We conclude that the privacy preservation through segmentation like CrowdMask [20] works well if the objective is to only preserve privacy without considering the cost of using the crowd to preserve privacy. But, if we introduce the trade-off between the cost to preserve privacy and the actual privacy preserved, we are limited by the amount of privacy preserved since the segment sizes may become larger in budget-constrained scenarios.

5.4 Experiment 2: Machine Learning Only

In this experiment, we measure how effective are machine learning models to detect private information in images?

For this experiment, we use pre-trained models for object detection[1], scene recognition [49] and text recognition [50, 37] to measure the amount of private information



Does the image contain any private information?

○YES ●NO

(a) NO, private visual cues present in image segment

Does the image contain any private information?

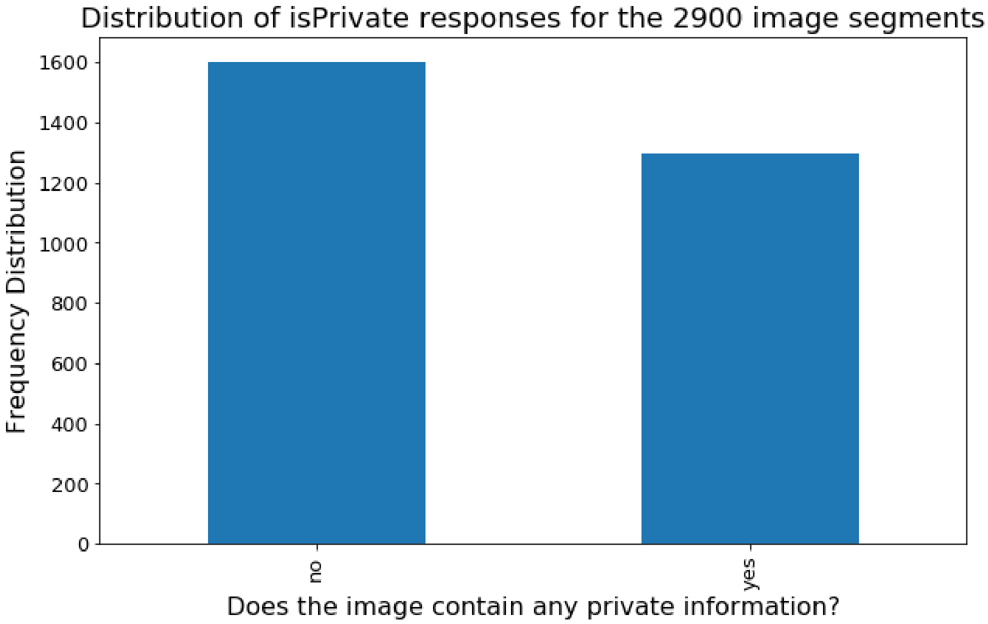
●YES ○NO

If yes, please select the appropriate category of private information?

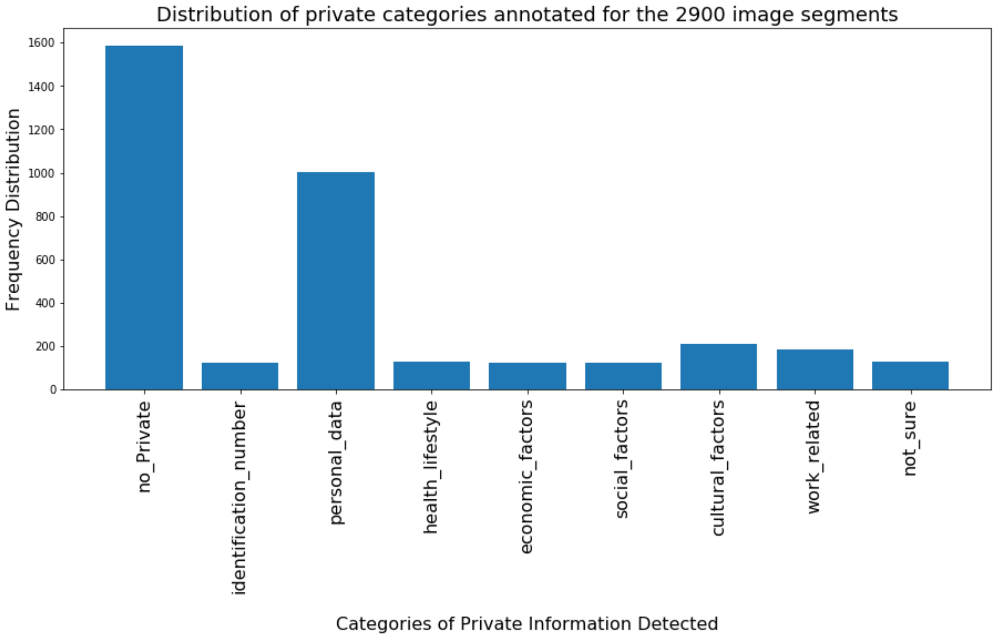
□ Identification Number - license plates, numbers on passport, visa, credit cards, ID cards, etc.
☑ Personal Data - person's face, gender, ethnicity (skin color), name, address, contact details, educational records, etc.
□ Health and Lifestyle - biometric information, physical disabilities, medical prescription, medicines, clinical text, etc.
□ Social Factors - political or trade memberships (flyers, books, magazines, badges), religious objects, religious symbols, etc.
□ Cultural Factors - ethicity, festivals, religious customs, traditions, ethnic clothing and traditions, etc.
□ Economic Factors - income, financial status, lifestyle, standard of living, etc.
□ Work-related - confidential data, specific informations, employee information, etc.
☑ Not Sure about the category

(b) Private visual cues present in the image segment

Figure 5.3: Answering tasks using our custom task interface for detecting private visual cues in the image



(a) Distribution of the presence of private information



(b) Distribution of annotated categories of private information

Figure 5.4: Results of crowdsourcing only experiment for detecting private information for 2900 image segments

that could be detected with off-the-shelf machine learning models. We use pre-trained machine learning models to simulate the scenario where the models do not have an annotated dataset for private information. Off-the-shelf machine learning models can detect private information like identity, skin color, ethnicity, gender, age, physical disabilities, or medical condition of an individual through the person object class with high accuracy. The text detection model can to some extent recognize keyword like passport but the overall reliability of private keywords is very low.

Experiment	only with Machine Labels		combining Machine Labels with Rules	
	Precision	Recall	Precision	Recall
person_body	1.00	0.985	1.00	0.985
person_face	1.00	1.00	1.00	1.00
person_nudity	0.606	1.00	0.606	1.00
physical_disabilities	0.757	1.00	0.757	1.00
clothing	0.00	0.00	1.00	1.00
license_plate	0.00	0.00	1.00	0.778
address_text	0.00	0.00	1.00	0.143
tickets	0.00	0.00	1.00	0.714
student_ID	0.00	0.00	1.00	1.00
datetime	0.00	0.00	1.00	0.40
medicines	0.00	0.00	1.00	0.667
credit_card	0.00	0.00	1.00	1.00
passport	0.00	0.00	1.00	0.833
drivers_license	0.00	0.00	0.00	0.00

Table 5.1: Experiments of preserving privacy by machine learning only

In Table 5.1, we quantitatively measure the amount of private information detected using off-the-shelf machine learning models. We define the quantitative measures of precision and recall for this experiment as,

$$Precision = \frac{\text{Number of correctly detected private visual cues}}{\text{Total number of private visual cues detected}} \tag{5.1}$$

$$Recall = \frac{\text{Number of correctly detected private visual cues}}{\text{Total number of private visual cues in the ground truth}} \tag{5.2}$$

Based on the results of using labels generated by off-the-shelf machine learning models, we wanted to check if combining the different machine-generated labels can increase the number of private information detected in the image. We create detection rules for private information using logical combinations of machine-generated labels. The rules map the machine-generated labels to the parameters humans apply to classify private visual cues. For example, consider the visual cue of a *passport*, a human would first classify it as a *book* object and then based on the text appearing in the visual cue classifies it as a passport if certain keywords like a country's name, the word "passport" is present in the detected book. To translate this process, as a rule, using on machine-generated labels, we first need to collect a list of country names and store it in the memory. Then we run the machine learning models to detect objects and recognize text on the image. Thus, the rule to detect a passport will be, " if detected_object is a book and "passport" exists in the detected_text or detected_text contains the name of a country". We repeated this process of derive the rules to detect the following private visual cues in the image: `person_body`, `person_face`, `person_nudity`, `physical_disabilities`, `clothing`, `license_plate`, `address_text`, `tickets`, `student_ID`, `medicines`, `credit_card` and `drivers_license`. The visual cues related to the *person* object class was the most straight-forward rule but visual cues like `address_all`, `tickets` which has a higher ratio of text to objects detected have longer, complex detection rules. We were unable to separate certain private visual cues containing a person's name since it is difficult to find an extensive collection of names. From Table 5.1, we find that the number of private information detected in the image increases using the detection rules. However, the number of private information detected using this rule-based approach depends on how the private information is annotated in the ground truth dataset and the machine-generated labels. The precision and recall computed for the private information category are comparatively higher than [31]. This may be because we used ground truth categories to create detection rules.

We want to measure the amount of private information that we can detect using the taxonomy of private information. The taxonomy of private information will be able to generalize the detection of private information.

In this variant of the machine learning-only experiment, we found that using the taxonomy to detect private information resulted in a high false-positive rate. When we analyzed the results, we found that the taxonomy was over-estimating the number of private information present in the image. For example, if the image of a person working out in the gym, the taxonomy correctly detects the person but flags additional private information for the image like nudity and physical disabilities. Thus, every time the taxonomy detects a person, it also states that the image exposes the nudity and physical disability attributes of the person. The reason for the high false-positive rates is attributed to the semantic similarity between the visual cues that contain private information. For example, the object instance of a person can disclose private information like age, gender, ethnicity, physical disabilities, nudity.

From this experiment, we conclude that an annotated dataset for private information is an absolute necessity for detecting private information using machine learning-based privacy preservation approaches. [31] proposed a model which could detect different private information in images, but the model used around a dataset of around 8,000 images that were manually, hand-annotated for private information. We also observe that the amount of private information detected using machine learning models is limited to the number of classes of private information annotated in the dataset.

5.5 Experiment 3: Context-based segmentation

In this experiment, we measure the amount of private information that is detected using our proposed context-based segmentation algorithm. We evaluate the performance of our algorithm in a budget-constrained scenario where only a limited budget is available for preserving privacy in the image. The budget-constrained scenario also serves as a good experimental condition to compare the context-based segmentation algorithm with the current crowdsourcing task segmentation approaches. In this constrained setting, we can get a better understanding of how effective our algorithm is to create crowdsourcing tasks to detect private information, such that the crowdsourcing tasks reduce the number of private information visible to the crowd.

We perform this experiment in two settings where we first vary the privacy threshold for a fixed obfuscation threshold. The context-segmentation determines the number of image segments created based on the privacy threshold. In this experimental setting, we evaluate how the number of crowdsourcing tasks created is affected by the computed privacy disclosure and privacy threshold for the image. From Figure 5.5, we observe that close to 33% of the image segments produced has 4 image segments. However, we have to bear in mind, that while the number of segments is the same as the number of large-segments (represented as the red colored vertical line) are not of the same size. We also note that there is close to 25% of the image segments which have 0 image segments which indicate that it is possible to detect private information in images using only the context-based segmentation algorithm without creating crowdsourcing tasks. By segmenting the image using image context, we have significantly reduced the number of image segments used in crowdsourcing tasks to detect private information, in comparison to current approaches like [20]. Thus, our proposed context-based segmentation algorithm is successful in minimizing the cost spent on detecting private information through crowdsourcing tasks.

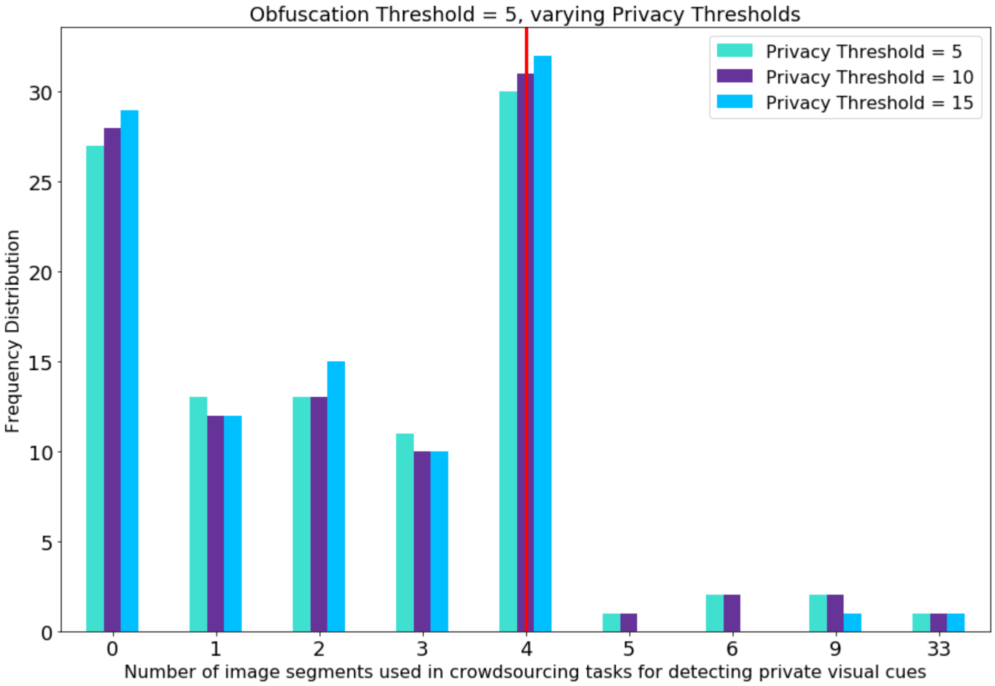


Figure 5.5: Distribution of the number of image segments created for crowdsourcing tasks for detecting the private information in the image by varying the privacy threshold

In our second experimental setting, we are interested in studying how the obfuscation threshold impacts the number of crowdsourcing tasks created. The obfuscation threshold refers to the permissible amount of black pixel for one image segment and ranges from 0 - 2. In this experiment, we vary the obfuscation threshold from low (0.5) to high (1.5) for a low privacy threshold of (5). The context-based segmentation by default does not send image segments that are completely obfuscated, where the obfuscation value would be 2. Even in this experiment, we have to bear in mind, that while the number of segments is the same as the number of large-segments (represented as the red colored vertical line) and the number of medium-segments (represented as the orange colored vertical line), are not comparable. From Figure 5.6, we observe considerable variations int he number of segments, but are unable to find conclusive reasons. Thus, we conclude that creating crowdsourcing segments based on privacy disclosure is the best suited for minimizing the cost for crowdsourcing. We also observe that a higher obfuscation threshold can maximize the amount of privacy preserved as observed from the responses of the crowd for the presence of private information in the image segment shown in Figure 5.7(b).

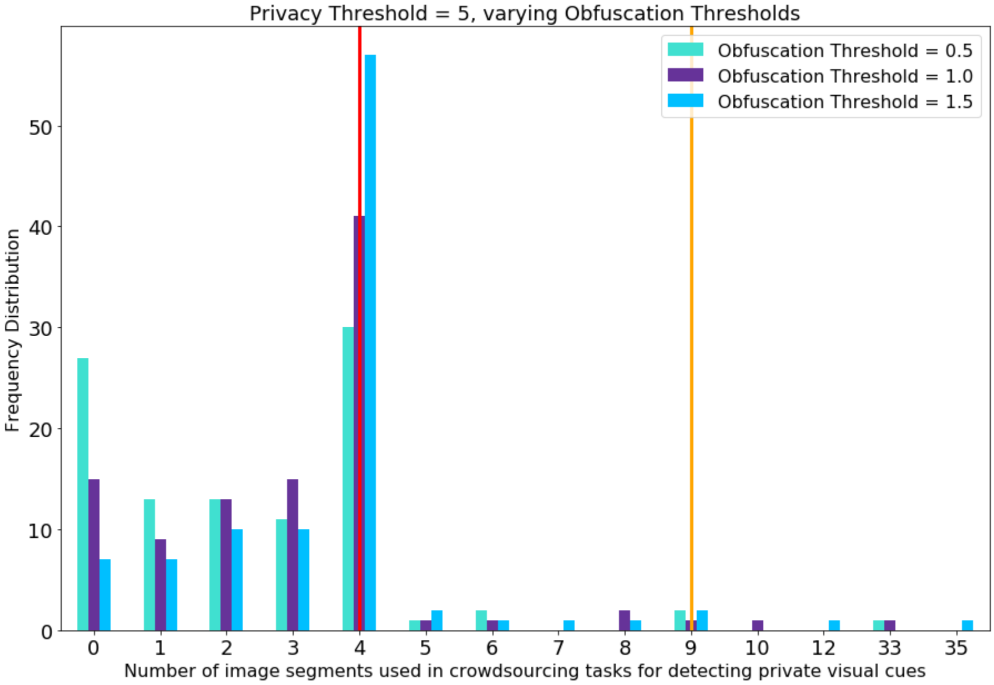
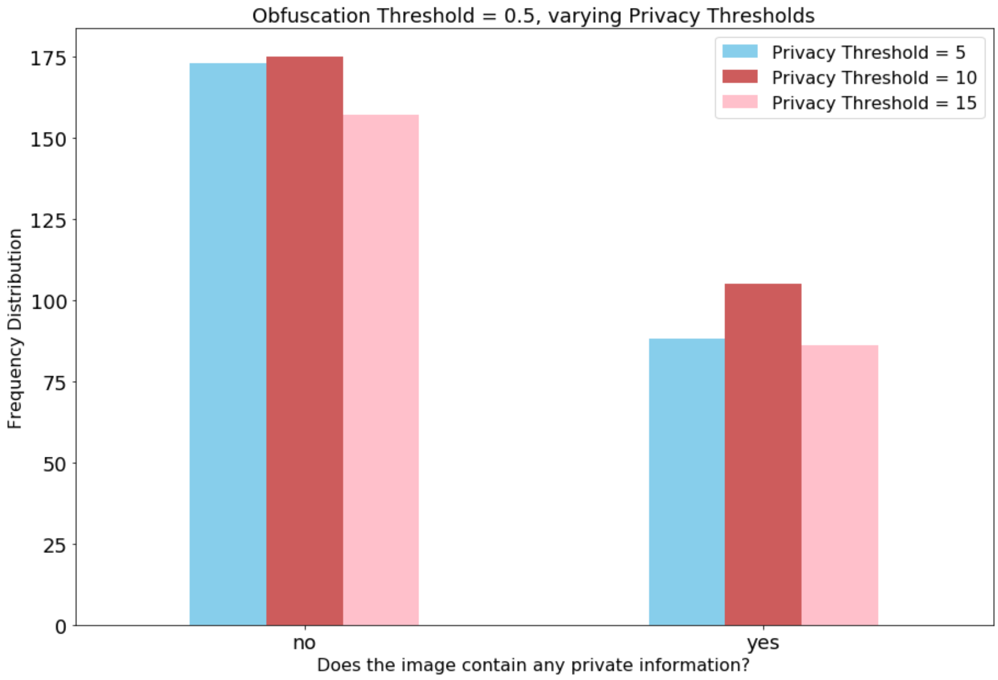
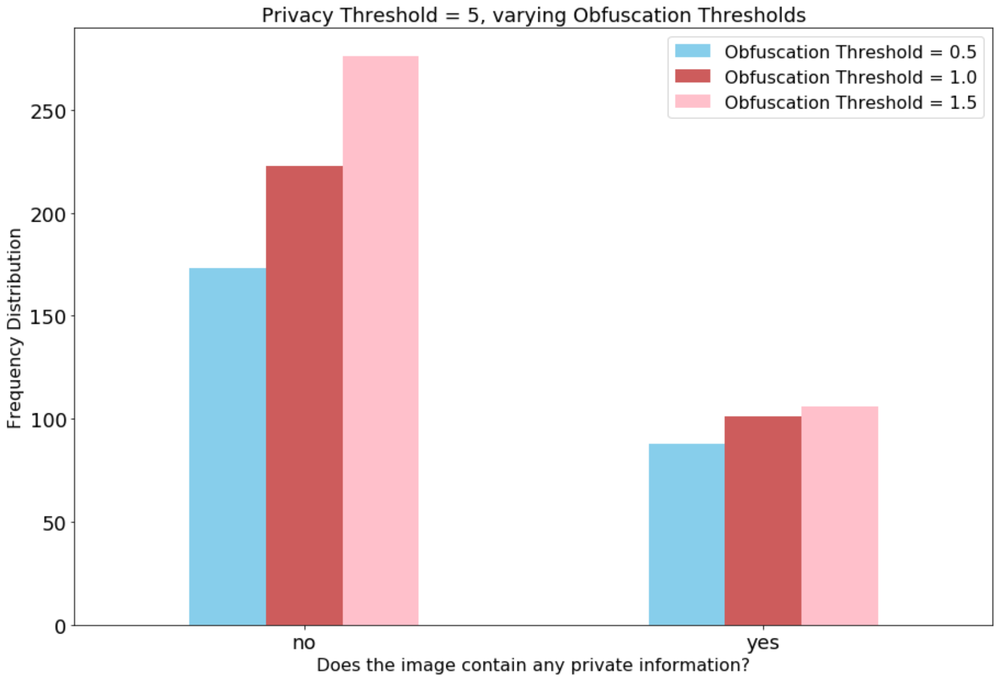


Figure 5.6: Distribution of the number of image segments created for crowdsourcing tasks for detecting the private information in the image by varying the obfuscation threshold

Finally, we are interested to understand the responses of the crowd on the categories of private information present in the image segments. From Figure 5.8(a), we find that on average a higher privacy threshold reduces the number of image segments, which supports the claim that the privacy disclosure-based segmentation approach. We observe that the `personal_data` category has the most responses, which indicates that a slightly fine-grained category list is required for crowdsourcing tasks to detect private information. Figure 5.8(b) represents no change in the categorization of private information present in the image however overall, the number of image segments in



(a) Varying privacy threshold values



(b) Varying obfuscation threshold values

Figure 5.7: Distribution of the responses from the crowd on whether the image contains any private information in the image segments

https://www.overleaf.com/project/5bdb31f5a6c4204079b0f752each category by varying the obfuscation thresholds is comparable to the number of segments created by varying the privacy threshold.

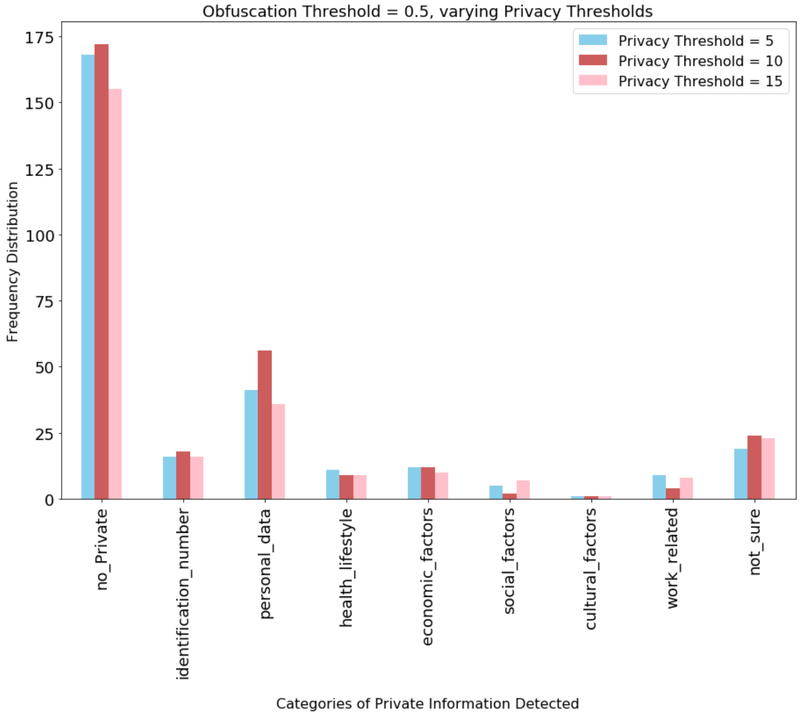
5.6 Discussion

In this chapter, we experimentally prove that our proposed context-based segmentation algorithm is effective in minimizing the cost using the crowd to detect private information in the image in comparison to current privacy-preserving task content segmentation approaches in crowdsourcing.

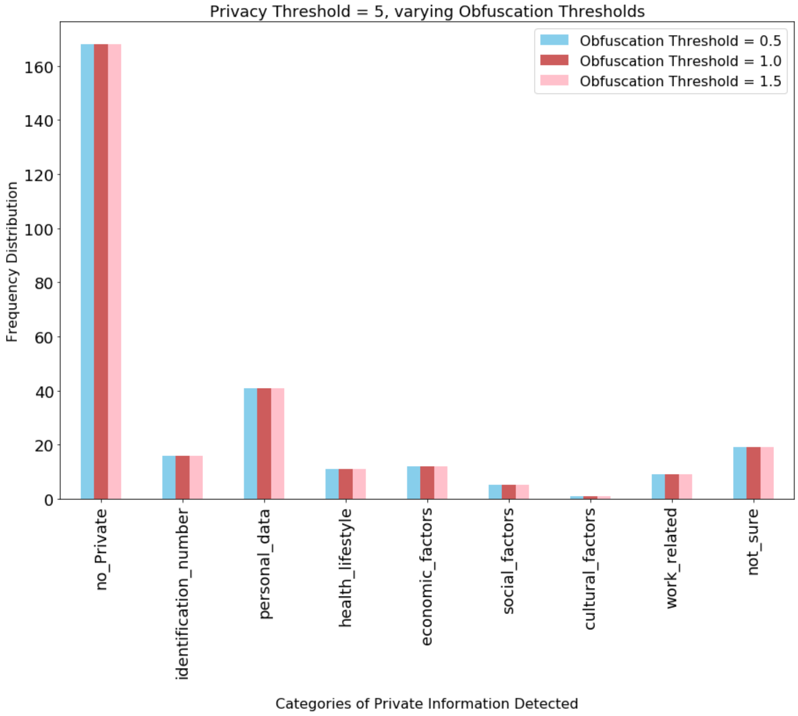
From the crowdsourcing-only experiment and the detection tasks created through the context-based segmentation algorithm, we note that the task design is not effective in getting responses from the crowd. This means, that we are unable to leverage the wisdom of the crowd for detecting private information. This means that a more granular definition of categories of private information is needed in the detection task. We also need to explicitly mention the different private information that belongs to each category of private information, since we found that the crowd mixed up private information belonging to the `social_factors` and `cultural_factors`. We recommend a more granular definition of private information categories so that we could reduce the large overlap of private information between the different categories of private information. We also recommend an aggregation method which penalizes responses that contribute to a large overlap of private information between categories. This aggregation method could reduce the number of inconsistent responses where the crowd mix-up the categories of private information.

From the context-based segmentation algorithm, we expect the amount of private information extracted by the crowd should be reduced, but the threat of information extraction in the detection crowdsourcing tasks still exists as an external threat to the crowdsourcing workflow used in our approach. We observe that a hybrid approach based on the disclosure score of private information in the image is effective to create privacy-preserved image segments. The privacy disclosure score is largely determined by the disclosure of private information by the object detection model (R_{Object}). This could be due to the skewness of the `VISPR` dataset to the `person_body, person_face` class labels which correspond to the 'person' object class, that is detected with high accuracy.

We found that our taxonomy of private information yields a large number of false positives due to the high semantic overlap between different private information classified to the same machine-generated label, for example, the 'person' object class maps to private information like the individual's identity, gender, ethnicity, physical disability and many more. This is an internal threat which arises due to the machine-generated labels, we could mitigate this threat by adding more models to get descriptive labels for each private information.



(a) Varying privacy threshold values



(b) Varying obfuscation threshold values

Figure 5.8: Distribution of the responses from the crowd on the categories of private information present in the image segments

Chapter 6

Conclusion

This thesis builds the foundation for future work which combines machine learning and crowdsourcing for preserving privacy in images. Our proposed hybrid workflow called context-based segmentation can be used for preserving privacy in images used in image analysis crowdsourcing tasks as well as for other application which uses images.

6.1 Contributions

The main contributions of this thesis are:

1. A comprehensive, **systematic literature study** on the current state of the art approaches of preserving privacy in images through machine learning and crowdsourcing. From the literature study, we find that privacy preservation methods in crowdsourcing are susceptible to workers extracting private information from the task, and become expensive at scale. Privacy preservation through machine learning show promising results but all the models were trained on an annotated dataset. Most of the datasets used for training machine learning models to preserve privacy were hand-annotated datasets which required a significant amount of time and effort. We identify a vicious loop between machine learning and crowdsourcing methods for privacy preservation. The research gap lies in finding a method to leverage the advantages of both machine learning and crowdsourcing to break-free from this vicious loop.

2. We have developed a **mapping of private information to visual cues** in images. This mapping can be used in different contexts and applications to detect private information. Our objective was to create a mapping which covers a broad set of private information defined in policies like the GDPR and map the private information to their corresponding visual cues in the image. This mapping is used in the context-based segmentation to detect the different private information which could visual cues or textual data in images. This mapping covers over 120 visual cues that contain private information in images, making it a one-of-its-kind classification of private information for images.

3. Our proposed **context-based segmentation** algorithm uses machine-generated labels to create privacy-preserving image segments which can be sent to the

crowd for verifying if there are additional private visual cues present in the image. We performed two experiments to evaluate how effective are machine learning and crowdsourcing methods in detecting private visual cues in the image. We perform another experiment on the proposed context-based segmentation approach to study the effectiveness of a hybrid human-machine approach in the detection of private visual cues in images. From these experiments, we conclude that machine learning models need an annotated dataset for private information for them to be effective in detecting private information, crowdsourcing approaches are expensive for detecting private information over the 100 images we used for experimentation. The context-based segmentation proved to successfully reduce the amount of private information disclosed in the image by creating privacy-preserving image segments. We also observe that leveraging the wisdom of the crowd to detect private information depends on the task design. In our case, we were able to collect responses which broadly indicated the category of private information but we were unable to use the collected responses to improve the mapping of private information to visual cues.

6.2 Reflection

We developed one of the first hybrid workflows for preserving privacy in images. This thesis creates a platform for future work on this topic. The context-based segmentation algorithm answers the main research of question of this thesis to detect and obfuscation private information by combining machine learning and crowdsourcing such that it minimizes the cost to detect private information in the image, and maximizes the amount of private information detected. We spent a significant amount of time designing our proposed workflow which has been made modular, to support future works like adding new machine learning models and improving taxonomy of private information present in the image. Answering our main research question has given us the confidence to become bold and add more components to improve the workflow, spend more time experimenting different configurations to assess the potential of the context-based segmentation algorithm but it requires more time which we did not have.

6.3 Future work

There is a lot of scope for future work. However, we recommend incorporating better machine learning models for generating machine-generated labels, an empirical study on different obfuscation methods for privacy preservation, and measuring the usefulness of the image after privacy-preservation through the context-based segmentation algorithm. We also suggest creating a taxonomy of private information which has fine-grained categories of private information and incorporates privacy policies beyond the GDPR. From a policy perspective, risk and quality assurances on the quality and quantity of privacy preserved in images is also a possibility

Bibliography

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. `https://github.com/matterport/Mask_RCNN`, 2017.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the 33rd AAAI conference on Artificial Intelligence. AAAI*, 2019.

[3] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual 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*, pages 333–342. ACM, 2010.

[4] L Elisa Celis, Sai Praneeth Reddy, Ishaan Preet Singh, and Shailesh Vaya. Assignment techniques for crowdsourcing sensitive tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 836–847. ACM, 2016.

[5] Jiawei Chen, Janusz Konrad, and Prakash 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*, pages 1570–1579, 2018.

[6] Kuang Chen, Akshay Kannan, Yoriyasu Yano, Joseph M Hellerstein, and Tapan S Parikh. Shreddr: pipelined paper digitization for low-resource organizations. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, page 3. ACM, 2012.

[7] Sen-Ching Samson Cheung, Herb Wildfeuer, Mehdi Nikkhah, Xiaoqing Zhu, and Waitian Tan. Learning sensitive images using generative models. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4128–4132. IEEE, 2018.

[8] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. Anonymizing k-facial attributes via adversarial perturbations. In *Proceedings of the 27th Inter-*

national Joint Conference on Artificial Intelligence, pages 656–662. AAAI Press, 2018.

[9] Richard Chow, Philippe Golle, and Jessica Staddon. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 893–901. ACM, 2008.

[10] Liting Du, Chenxi Xia, Zhaohua Deng, Gary Lu, Shuxu Xia, and Jingdong Ma. A machine learning based approach to identify protected health information in chinese clinical text. *International journal of medical informatics*, 116:24–32, 2018.

[11] Liyue Fan. Image pixelization with differential privacy. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 148–162. Springer, 2018.

[12] Anhong Guo, Anuraag Jain, Shomiron Ghose, Gierad Laput, Chris Harrison, and Jeffrey P Bigham. Crowd-ai camera sensing in the real world. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):111, 2018.

[13] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey 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*, pages 939–948, 2019.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] Zhida Huang, Zhuoyao Zhong, Lei Sun, and Qiang Huo. Mask r-cnn with pyramid attention network for scene text detection. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 764–772. IEEE, 2019.

[16] Jinyuan Jia and Neil Zhenqiang Gong. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 513–529, 2018.

[17] Hiroshi Kajino, Hiromi Arai, and Hisashi Kashima. Preserving worker privacy in crowdsourcing. *Data Mining and Knowledge Discovery*, 28(5-6):1314–1335, 2014.

[18] Hiroshi Kajino, Yukino Baba, and Hisashi Kashima. Instance-privacy preserving crowdsourcing. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[19] Ece Kamar, Ashish Kapoor, and Eric Horvitz. Lifelong learning for acquiring the wisdom of the crowd. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[20] Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey P Bigham, Jaime Teevan, Ece Kamar, and Walter S Lasecki. Crowdmask: Using crowds to preserve privacy in crowd-powered systems via progressive filtering. In *Proceedings of the AAAI Conference on Human Computation (HCOMP 2017)., HCOMP*, volume 17, 2017.

[21] Evgeny Krivosheev, Fabio Casati, Marcos Baez, and Boualem Benatallah. Combining crowd and machines for multi-predicate item screening. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):97, 2018.

[22] Gierad Laput, Walter S Lasecki, Jason Wiese, Robert Xiao, Jeffrey P Bigham, and Chris Harrison. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1935–1944. ACM, 2015.

[23] Walter S Lasecki, Mitchell Gordon, Winnie Leung, Ellen Lim, Jeffrey P Bigham, and Steven 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*, pages 1945–1954. ACM, 2015.

[24] Walter S Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P Bigham. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1203–1212. ACM, 2013.

[25] Walter S Lasecki, Jaime Teevan, and Ece Kamar. Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 248–256. ACM, 2014.

[26] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1343–1351. IEEE, 2017.

[27] Greg Little and Yu-An Sun. Human ocr: Insights from a complex human computation process. In *Workshop on Crowdsourcing and Human Computation, Services, Studies and Platforms, ACM CHI*. Citeseer, 2011.

[28] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.

[29] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2774–2783. IEEE, 2017.

[30] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[31] Tribhuvanesh Orekondy, Mario Fritz, and Bernt 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*, pages 8466–8475, 2018.

[32] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3686–3695, 2017.

[33] Ismini Psychoula, Erinc Merdivan, Deepika Singh, Liming Chen, Feng Chen, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. A deep learning approach for privacy preservation in assisted living. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 710–715. IEEE, 2018.

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[35] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131–151, 2016.

[36] David Sánchez and Montserrat Batet. Toward sensitive document release with privacy guarantees. *Engineering Applications of Artificial Intelligence*, 59:23–34, 2017.

[37] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

[38] Xuemeng Song, Xiang Wang, Liqiang Nie, Xiangnan He, Zhumin Chen, and Wei Liu. A personal privacy preserving framework: I let you know who can see what. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 295–304. ACM, 2018.

[39] Haipei Sun, Boxiang Dong, Bo Zhang, Wendy Hui Wang, and Murat Kantarcioglu. Sensitive task assignments in crowdsourcing markets with colluding workers. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 377–388. IEEE, 2018.

[40] Saiganesh Swaminathan, Raymond Fok, Fanglin Chen, Ting-Hao Kenneth Huang, Irene Lin, Rohan Jadvani, Walter S Lasecki, and Jeffrey P Bigham. Wearmail: On-the-go access to information in your email with a privacy-preserving human computation workflow. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 807–815. ACM, 2017.

[41] Lav R Varshney. Privacy and reliability in crowdsourcing service delivery. In *SRII Global Conference (SRII), 2012 Annual*, pages 55–60. IEEE, 2012.

[42] Lav R Varshney, Aditya Vempaty, and Pramod K Varshney. Assuring privacy and reliability in crowdsourcing with coding. In *Information Theory and Applications Workshop (ITA), 2014*, pages 1–6. IEEE, 2014.

[43] Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18:193–1, 2017.

[44] Idalides J Vergara-Laurens, Luis G Jaimes, and Miguel A Labrador. Privacy-preserving mechanisms for crowdsensing: Survey and research challenges. *IEEE Internet of Things Journal*, 4(4):855–869, 2017.

[45] Li Wang, Jun Jie Shi, Chen Chen, and Sheng Zhong. Privacy-preserving face detection based on linear and nonlinear kernels. *Multimedia Tools and Applications*, 77(6):7261–7281, 2018.

[46] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 606–624, 2018.

[47] Zhiqiang Yang, Sheng Zhong, and Rebecca N Wright. Privacy-preserving classification of customer data without loss of accuracy. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 92–102. SIAM, 2005.

[48] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338. IEEE, 2011.

[49] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[50] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.