Privacy-Preserving Face Redaction Using Crowdsourcing

Abdullah B. Alshaibani,[1] Sylvia T. Carrell,[2] Li-Hsin Tseng, Jungmin Shin,[3] Alexander J. Quinn[1]
[1]Purdue University [2]Sandia National Laboratories [3]Korea Military Academy
{aalshai, aq}@purdue.edu, scarrel@sandia.gov, lizzietseng@google.com, c16267@mnd.go.kr

Abstract

Redaction of private information from images is the kind of tedious, yet context-independent, task for which crowdsourcing is especially well suited. Despite tremendous progress, machine learning is not keeping pace with the needs of sensitive applications in which inadvertent disclosure could have real-world consequences. Human workers can detect faces that machines cannot; however, an open call to crowds would entail disclosure. We present IntoFocus, a method for engaging crowd workers to redact faces from images without disclosing the facial identities of people depicted. The method works iteratively, starting with a heavily filtered form of the image, and gradually reducing the strength of the filter, with a different set of workers reviewing the image at each step. IntoFocus exploits the gap between the filter level at which a face becomes unidentifiable and the level at which it becomes undetectable. To calibrate the algorithm, we performed a perceptual study of detection and identification of faces in images filtered with the median filter. We present the system design, the results of the perception study, and the results of a summative evaluation of the system.

Introduction

Robust preservation of privacy is one of the key barriers to a future where digital work can be transferred to crowd workers with the ease and confidence with which we send files today (Kittur et al. 2013). Examples of tasks that should maintain privacy are:

1. Create derivative works from photos taken by children on a class field trip.
2. Search images from private social media accounts for evidence of violence or bullying.
3. Redact private information from court records released for public disclosure laws.

Very accurate detection of faces is a necessity to minimize the potential for real harm due to accidental disclosure of identities. Despite the impressive progress made with automated face detection in recent years, challenges such as occlusions, pose, illumination (low or high), atypical skin tones, skin-colored backgrounds, and weather (rain, snow, haze) (Nada et al. 2018; Buolamwini and Gebru 2018) still give human perception the edge over machines.

Even a modest goal of 95% recall (proportion of faces detected) with 95% precision (proportion of matches that are actually faces) remains beyond the reach of any current algorithm that we are aware of (Jain and Learned-Miller 2010; Zhu et al. 2017; Nada et al. 2018; Zhang et al. 2020). Of course, applications with the potential for real harm in case of disclosure demand much better than 95%.

This research leverages crowd workers efficiently—not because humans will always be more accurate than machines for face redaction, but because combining crowd workers with machines has the potential to exceed the capabilities of both. Although we focus exclusively on human perception in this paper, an ultimate solution would combine them.

We present *IntoFocus*, a method and system that engages crowd workers to redact faces from still images. It starts by showing workers a heavily filtered form of the image and asking them to add ellipses to regions that may contain the specified type of information. Successive iterations present slightly less-filtered images while blocking regions marked as potentially sensitive in prior iterations.

The primary contributions of this paper are as follows:

1. The IntoFocus method allows consistent privacy-preserving face redaction by crowd workers.
2. We present a study of face detection and identification *by humans* in median-filtered images.
3. Our evaluation validated that the IntoFocus method results effectively detects faces without exposing crowd workers to facial identities of depicted persons.

Related Work

The foundations of this work can be understood in terms of ① motivating applications, ② privacy-preserving crowdsourcing, ③ design and technical foundations, and ④ human perception of faces.

Motivating Applications

Crowdsourcing and human computation began to gain prominence in 2005 in research (Quinn and Bederson 2011) and with the founding of influential commercial services, such as Mechanical Turk. Initial applications were limited to data that could be shared publicly. One of the first

published references of the need for privacy—for the requester's data—was in 2010, in relation to document processing workflows (Karnin, Walach, and Drory 2010).

The risk to humans became palpable with VizWiz, a mobile application that allows blind people to get help with everyday situations by sending a photo and a spoken question to workers on Amazon Mechanical Turk (AMT) (Bigham et al. 2010). VizWiz holds the risk of sharing sensitive information inadvertently included in the picture. A similar dilemma exists when robots are assisted by crowd workers. Sorokin used such an approach to enable robots to grasp unfamiliar objects (Sorokin et al. 2010). The robot sends images of the object to workers, who then draw contours to help the robot grasp the object. The IntoFocus *method* could someday be integrated into such systems to enable robust redaction of sensitive content prior to presenting the image to the workers who will render the assistance.

Privacy-Preserving Crowdsourcing

WearMail (Swaminathan et al. 2017) introduced the use of a system that allows workers to search through a person's email to answer a specific question that they have. They implemented privacy mechanisms that allow the requester to blacklist specific words and hide them from the workers. Work by Deng et al. (Deng, Krause, and Fei-Fei 2013) used blurred images of birds in experiments where workers were allowed to reveal small regions of the image that would help them accurately categorize the type of bird they saw, without revealing the entire image. Similar to IntoFocus, these methods reveal small amounts of information to preserve the privacy of the subjects.

Lasecki and collaborators have been highly active in developing methods for privacy-preserving crowdsourcing applications. One application engages crowd workers for behavioral coding of video (e.g., for social science research) (Lasecki et al. 2013). Of their many deveoloped crowdsourcing systems, CrowdMask is the most similar to IntoFocus, with respect to the purpose.

CrowdMask (Lasecki et al. 2015a; Kaur et al. 2017) segments a single image into smaller segments and asks the workers to annotate segments that contain sensitive information or that may be adjacent to sensitive information. It uses a pyramid workflow, which is effective for tasks where judgment about a particular segment can be made based on local information. However, because workers do not see the full photo, they might not be able to judge if a specific region contains private information (e.g., because it is cut in half or is otherwise taken out of context) and it does not account for the risk of having all the information in a single segment. In contrast, IntoFocus shows the entire image, but uses gradual revelation to ensure that sensitive regions are not disclosed. In a follow-up, Lasecki et al. used Gaussian blur in a single layer and documented that behaviors can still be identified even when a video is blurred sufficiently to hide identities (Lasecki et al. 2015b).

Lasecki et al. (Lasecki, Teevan, and Kamar 2014) have demonstrated the risks of completely trusting crowd workers with sensitive information. They showed that when some workers were given incentives, they would sabotage a task. They also showed that there are other workers that would not let such things happen, who went out of their way to report what happened. A few recent efforts have proposed methods for addressing this challenge for image-oriented tasks.

One of the first such efforts involved a protocol for instance-privacy based on clipping regions. It used a clipping function based on additional feedback provided by the requester (Kajino, Baba, and Kashima 2014). The need for requester involvement was a limitation, and its "instance-clipping protocol" was not a comprehensive solution.

Design and Technical Foundations

Peekaboom (von Ahn, Liu, and Blum 2006) introduced the combination of crowd workers and object detection and identification. They used two workers: one tasked to reveal portions of an image and another tasked to identify what is in the image. The work shows that even with limited revelation, humans are still able to identify objects. With the revelation of specific regions, humans are able to find or identify the objects in the images. We use this information to build a system that, given a highly filtered image, slowly reveals safe regions in order to help humans in finding the regions we want to remain hidden.

Efforts to enhance image segmentation have included strategies that ask human workers to annotate objects in the foreground via a variety of interactions (Gurari et al. 2016). The issue with regards to privacy is whether the people who appear in the image are aware that their image has been taken. Our focus is to redact the faces in an image before submitting it to crowdsourcing platform to solve the task, whether the face is in the foreground or the background. Thus, the workers would have a redacted image of the object/subject being segmented instead of a clear image.

Human Perception

Lewis et al. (Lewis and Edmonds 2003) found that humans can find faces more quickly when bodies are attached. This shows that people search the entire scene when looking for faces.

While making sense of an image had the effect of enhancing people's face detection abilities, identifying people of different races had the opposite effect (Lindsay, Jack, and Christian 1991). Results show that recognition memory is better for faces of the same race as the participant. This shows that skin tone is a factor that needs to be considered in the face-perception experiments.

Das et al. (Das et al. 2017) showed that, when tasked to find specific objects in images, humans searched in regions that are different from where deep networks searched. Their research shows that humans have a better understanding of images and the physical world than deep networks. They also found that when deep networks are programmed to search in the humans' regions, they performed better. The same can be achieved for face detection algorithms, where humans have a lifetime of experience in detecting faces.

IntoFocus Method

To safely leverage crowd workers for face redaction, IntoFocus uses an algorithm based on progressive image clarifi-

cation (i.e., presenting the image to workers with decreasing filter levels).

The input is an image containing any number of faces at any scale. We assume that the scale of possible faces is unknown, and that machine detection might fail for some of them. To enable clear measurements of our technique, we do not integrate machine detection. The output will be the same image, but with all faces redacted.

The process proceeds iteratively:

Stage 1: In this first step, very large faces are redacted. The image is filtered using median filter with a $ksize$ (kernel size) adequate to render typical faces of any size unrecognizable to humans. At this level, only very large faces will be perceivable as faces. For 640×640 images, we start with a kernel size of 85×85. We will refer to this as a *filter level* of 85. (Our method for selecting these constants will be discussed later in this section.)

The heavily-filtered image is presented to workers, who are asked to annotate all regions of the image that contain any part of a face. They added ellipses on portions of the image that contain any part of a face.

To reduce the chance of disclosure of identifying information, we collect multiple judgments from independent workers. Potential sources of variation of the quality of work include: differences in individual perception abilities, inattention, laziness, and malicious subterfuge. We combine the judgments using a union; if any worker identifies a pixel as belonging to a face, then we record it as such. In our trials, we collected 3 judgments per filter level, but the number of judgments collected could be configured to suit the security and affordability requirements of a given application.

Some over-redaction (false positives) is possible. This is considered acceptable based on the premise for IntoFocus. In our target applications, protection of human identities is a higher priority than preservation of other content. Our implementation does not actively defend against deliberate over-redaction by malicious workers, but we believe it could be addressed by using heuristics based on worker behavior and/or low-level image characteristics (e.g., if a worker flagged a texture such as grass or sand, which is extremely unlikely to contain a face, then flag for review). An attention-check image is used to measure over-redaction, but the measure applies to image sets, not each specific image.

Stage 2: The original clear image is filtered using a median filter with a lower $ksize$ value, we use $ksize = 53$. Regions identified by *any* worker in stage 1 are redacted, by filtering with a *higher* $ksize$ value. In our implementation, we use $ksize = 113$ for regions marked as faces in stage 1. A second cohort of workers are presented with this image and asked to mark any faces that are perceptible at this $ksize$ value. The interface is the same as before.

Stage *i*+1: With each successive stage, the $ksize$ value is decreased, allowing smaller faces to be redacted by the workers. Regions marked as faces in stage $i+1$ are redacted by filtering with the $ksize$ value from stage i.

Stage *n*: The final stage uses a small $ksize$ value to deidentify the smallest faces that could otherwise be recognizable to a worker who was familiar with the depicted person. Our implementation uses a $ksize$ value of $ksize = 9$ in the final stage. For added protection against disclosure of very small faces, additional stages could be added.

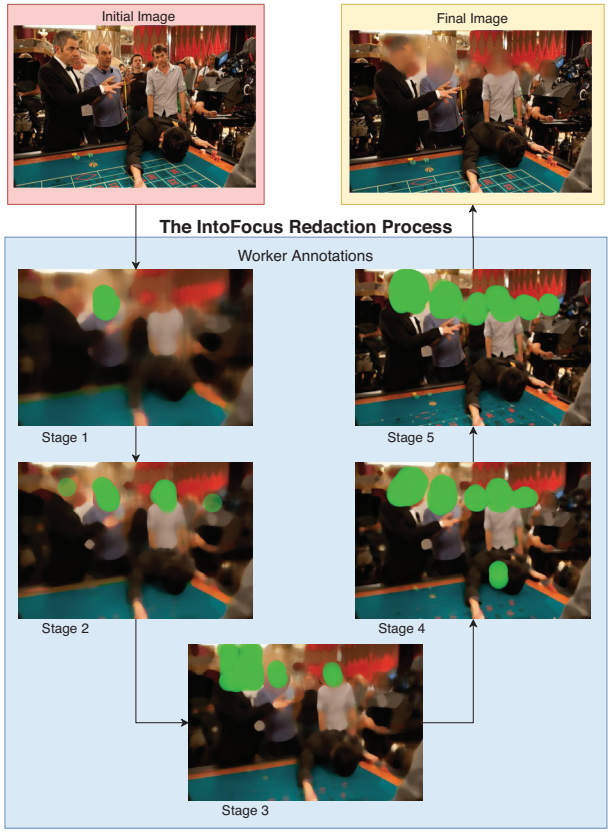

Figure 1: The diagram shows an example of the full flow of the system with 5 stages and a minimum of 3 workers for a single stage. The top left image is what the requester sent to be redacted. The green blobs are the workers' detections. The image at the top right is the resulting image from the redaction process.

Figure 1 shows the progress of an image as it goes through the IntoFocus method. The green blobs are the regions that were redacted by the workers at each stage. At the end of the final stage, the system would release an image that has the regions selected by the workers redacted and all other regions still visible. Now the image can be safely uploaded into a crowdsourcing platform to complete some other task without compromising the people in the image.

Preserving the privacy of people in an image does not end with hiding information in photos with regards to a single worker—it needs to hide the information from any and all workers. Varshney (Varshney 2012) explored what affects a worker's reliability, and that some workers would collaborate with others to extract the information they needed. Workers can target a task and try to extract information from it. The requester's goal is to hinder their progression and stop those attempts.

Parameters

This section describes some preliminary explorations that led to our choice of the median filter for the "filter" operation as well as a two-part face perception study to find the specific values of filter kernel size ($ksize$) used in the IntoFocus algorithm. The process for choosing these values is explained so that others may understand our design rationale and consider future improvements and optimizations.

Filter Method IntoFocus requires a filter operation that reveals enough fidelity to discern the outer contours of a face, while concealing smaller features that could be used to recognize the depicted person's identity (e.g., nose, eyes). We considered 5 filters: Gaussian, scatter, square pixelation, unfocus, and median. Gaussian and pixelation were eliminated due to known attacks that allow identification of text and/or faces from obfuscated images (Gross et al. 2006; Hardie, Barnard, and Armstrong 1997; Hummel, Kimia, and Zucker 1987). Unfocus was eliminated because it results in images that are qualitatively similar to Gaussian blur, and thus we suspect it may also be vulnerable to those attacks.

Scatter—a filter which displaces each pixel by a random distance and in a random direction—results in greater destruction of information due to its stochasticity. However, we found that when those images are subsequently processed with a Gaussian blur, the images become perceptually identical to Gaussian filtered images. In other words, some of the obfuscation effect is reversible. Consequently, we believe that more iterations of the IntoFocus algorithm would have been required had we selected the scatter filter.

Decision: We chose the median filter because it affords fewer opportunities for attack, and a median-filtered image cannot be further clarified. While we do not claim this choice to be optimal, our experience—including ad hoc exploration by the authors—indicates that it will reduce the number of iterations required to effectively redact an image, without disclosing the facial identities of persons depicted.

Parameters to Median Filter The median filter depends on a value, called the $ksize$. The filter creates a window of size $ksize\ (width) \times ksize\ (height)$ centered at each pixel, and computes the median of intensities of all pixels (for each color channel). The resulting median value for each channel becomes the new intensity for that pixel. In our implementation, we used the Python Pillow library's median function, which requires the $ksize$ to be an odd number. The filter also needs the dimension of all the images to be in the same range.

Face Perception Studies

To calculate thresholds, we conducted a two-part study of (a) face detection (Figure 2) and (b) face identification by humans (Figure 3).

The ultimate goal was to find optimal filter levels that would ensure that IntoFocus can detect each face with some probability (e.g., $Pr(anyworkerdetects) \geq 0.99$) while limiting the risk that any worker identifies a face to some low probability (e.g., $Pr(anyworkeridentifies) \leq 0.02$). To do this, we needed to answer a key question: *If a face can*

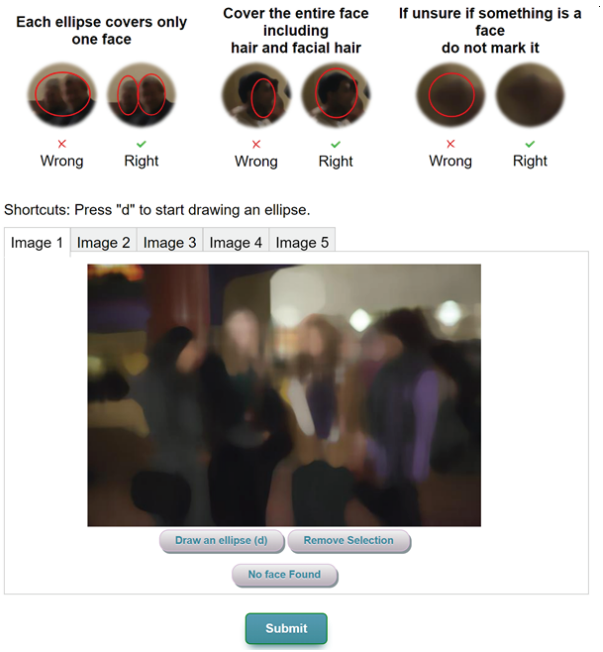

Figure 2: Face detection study. Participants annotated each face they detected with an ellipse. They could select 'No face found' button if unable to detect any faces.

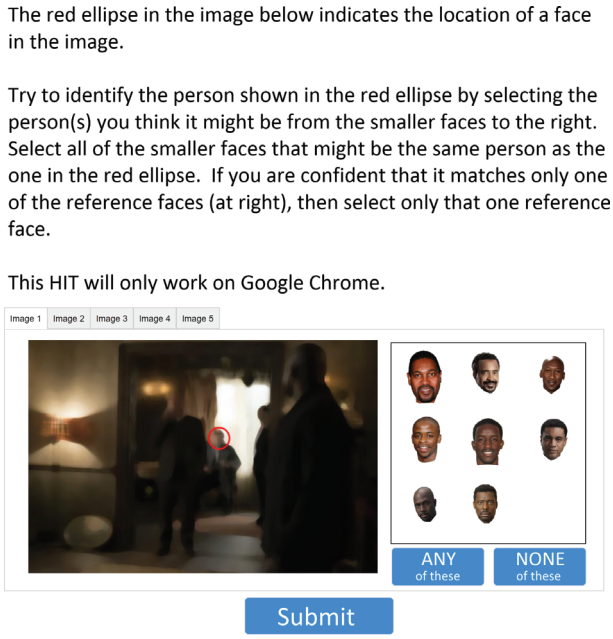

Figure 3: Face Identification study. Participants selected the reference face (right) matching the person in the main image (left) who was marked with a red ellipse. Multiple faces could be selected in case they were able to eliminate some reference faces from consideration, but could not identify the subject face (left) as definitely matching *one* of the reference faces (right).

be detected by N% of people at blur level k_{detect} (or lower), what is the minimum blur level at which no more than M% of people can identify the face, supposing they knew the person or had some reference photo available. We took $N = 98\%$ and $M = 2\%$.

Study Design Rationale The IntoFocus system relies heavily on the selection of filter levels that allow detection, but prevent identification, of faces. Without mapping the people's performance using the selected filter method, these filter levels are only random guesses. Thus, the first task is to show that there is a gap in filter levels between detection and identification. Then, using that information, we apply a model that extracts the filter levels to use.

Study Part 1: Detection The task is to add an ellipse on each of the faces that can be seen in the image. Ellipses can be modified and/or removed after they are added. In each HIT, one image served as an attention check. The image had a reduced blur level and all the faces could be easily detected. These images were to ensure the validity of the data and that each participant was performing the task correctly.

Study Part 2: Identification Participants were presented with a *main* image with one face outlined with an red ellipse. (We applied the ellipse.) They then attempted to match the depicted person to one of eight (8) reference faces.

Reference faces were selected with the same gender, hair color, and skin tone of the person in the ellipse. This was to minimize the chance that participants might guess correctly based on characteristics other than the face. (The scope of this research is limited to *facial* identities.)

Each image had only one red ellipse, even if other faces were present. This ensured consistency in our study design.

Dataset The dataset is a combination of two datasets. The first dataset was the IMDB-WIKI dataset (Rothe, Timofte, and Gool 2016), which was chosen because it provided many images from awards ceremonies, behind the scenes shots, and portrait images of all the actors were easily obtainable. There were people in both the foreground and the background of the images. In some images, the people were camouflaged or hidden in a corner of the image. The largest face was 400 pixels in width and 600 pixels in height.

The second dataset was of random people, posing for the camera, talking amongst themselves, eating, working on computers, cleaning, advertising, and other activities. The images contained people in the foreground and background of the image. The smallest face in the dataset was 3 pixels in width and the largest face was 200 pixels in width.

The reason behind building a dataset from scratch was to truly test how the system performs when the workers do not have prior knowledge of the images or when they were taken. These images were taken using a Google Pixel cellphone camera with HDR (high dynamic range). The dataset contained people of mixed ethnic backgrounds, and all were between the ages of 18 and 35. There was a total of 60 images used in the experiment. The datasets covered faces of all colors, shapes, and sizes. There was a total of 157 different faces to select from in this study. There was a total of 336 faces in the images.

Face Selection For the identification study, a set of 8 reference faces are presented next to each image. Workers are asked to select any faces that they believe are in the image. If they cannot match any of the reference faces to the main image, they can click a button labeled "don't know."

For any main image, we evaluate success with respect to only one of the depicted faces. (The study design becomes intractable if we try to evaluate success for all faces.) Therefore, in each trial, exactly one of the reference faces is present in the main image.

Success is indicated when workers choose "I don't know" or guess with random probability, based on the number of choices offered (i.e., 12.5% for eight reference images). We cannot judge success on any individual trial, but we can measure the rate of success for a group of trials.

Since this work is solely focused on *facial* identities, workers should not be able to narrow the set of reference faces by any characteristics other than the facial identity. Therefore, we used k-anonymity (Sweeney 2002) with $k = 8$ as our measure, using reference faces having the same non-facial characteristics: hair color, hair length, and skin tone.

Evaluation The study was initially designed to use in-person participants in a lab. Due to the COVID-19 pandemic, it was ultimately conducted online using workers on Mechanical Turk. To reduce redundant work (and control costs), we used a binary search to find the thresholds (filter level) where each face becomes detectable and identifiable. A pilot study was performed to find the filter level ranges for detection and identification for each of the images. In the study, each of the five images had a different filter level based on the ranges that were provided by the pilot study.

In the detection study, the search tree was used to find the filter level for each image where only one worker out of 25 is unable to detect the face. That filter level being searched for represents the point where the image starts becoming too filtered for people to detect faces. The starting point and boundaries used in the binary search algorithm were obtained from a pilot study on the same images in an in-lab study. The upper and lower bounds were the highest detection and the lowest detection filter levels, respectively. The starting point was the median filter level of all the participants in the pilot study.

In the identification study, the search tree was used to find the filter level where only one worker was able to correctly identify the face. That filter level being searched for represents the point where the image becomes too filtered for people to correctly identify the faces. The upper bound for the identification study was the lowest filter level that allowed detection. The lower bound was the filter level that allowed all the participants of the pilot study to correctly identify the person. If the face requires a filter level higher than the upper boundary, the filter level would be incremented by a value of 2 (the nearest odd number) until the filter level is found.

When the identification filter level is higher than the filter level where all the participants are able to detect, then we have the probability that a face is identified, given that the face is detected (in the identification study, the locations of the faces to be identified are given). Thus, the probability re-

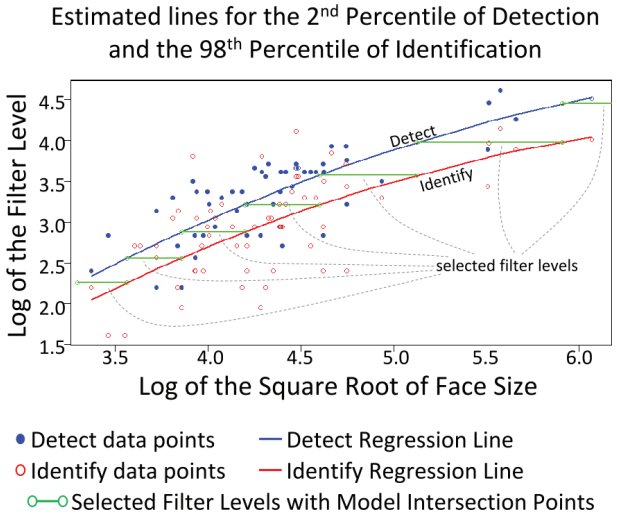

Figure 4: Study results. The blue line and points are the second percentile of the detect values (the point right before the detection rate reaches 100%). The red line and points are the 98th percentile of the identify values (the point right before the identification rate reaches 0%). The green line is the staircase model that was used to select the filter levels for the IntoFocus method.

quired is $P(identify \cap detection)$. The new probability of identification value can be calculated using the Kolmogorov axiom (Kolmogorov 1956), $P(identify \cap detection) = P(identify|detect)P(detect)$.

For each node in the search tree, 25 workers were hired to perform the given task (either detection or identification). Workers that did not pass the attention check were replaced. No workers were rejected in this study. After finding the filter levels for each image, the images were ordered by the size (width and height) of the faces, then a polynomial regression was performed to estimate the filter levels for detection and identification for different-sized faces (Figure 4).

The results were evaluated by taking the blur levels when each face was detected and identified. Using the 98th percentile for the identification and the 2nd percentile for the detection, we get a region where all faces $\pm 2\%$ are detectable and none of the faces $\pm 2\%$ are identifiable. Now that the boundary is set, starting from the smallest face size to be protected, a vertical line is drawn from the beginning of the identification line until it intersects with the detection line and at the intersection point, a horizontal line is drawn until it intersects with the identification line. That horizontal line represents the lowest filter level used. The process is repeated, creating a staircase, and each horizontal line found is a stage to be used in the IntoFocus method. Each task was offered for \$0.75, with a total of 2844 assignments (an assignment in AMT refers to the agreement between the person requesting the task and the person performing the task). The hourly rate was \$13.99. The total cost for mapping people's performance on 60 median filtered images was $\$2,986.20$.

Results The results show that there is a gap in the filter levels between the lowest detect and the highest identify (Figure 4). The model starts with a face width and height of 27 pixels, projects horizontally to the identification line, then projects vertically to the detection line. The horizontal projections were the filter levels used in the IntoFocus method. The model stops when the horizontal line no longer intersects with the identification line. The resulting filter levels were $(85, 53, 35, 25, 17, 13, 9)$. The results confirm our hypothesis: a gap exists between when participants are able to detect and when they are able to identify a face. The plot in figure 4 uses only the 2nd percentile for detection (the point where almost every person is able to detect the faces), and the 98th percentile for identification (the point where at most one person is able to identify). Even though the values are the extremes in both cases, the two are separable. Because the values are the extremes and (in the identification aspect of the study) the location of the face being identified was provided, we can see that in some cases, identification has a higher filter level than detection. These results are possible and were not adjusted for, because they represent a possible identification before 98% of the population can detect.

Decision: For these images (longest dimension = 640 pixels), our model gave us 7 iterations, with the following $ksize$ values: $(85, 53, 35, 25, 17, 13, 9)$. These numbers represent the $ksize$—i.e., width and height, in pixels, of the window—used for the median filter. Thus, in the first iteration, the image is filtered with $ksize = 85$. In the seventh (and last) iteration, the image is filtered with $ksize = 9$.

Initially, we planned to display the faces discovered by workers at the same $ksize$ value with which they were found. For example, faces discovered in stage 1 ($ksize = 85$) would be shown with $ksize = 85$ in all subsequent iterations. However, when shown in the context of an image that was filtered at a lower level (e.g., $ksize = 35$), we found that the faces were easier to recognize. We considered concealing them entirely (i.e., solid black), but that might impede discovery of other faces in future iterations. Therefore, we opted to filter the faces at a higher $ksize$ value.

Decision: After each iteration, the faces discovered by workers are further filtered with the following $ksize$ values: $113, 85, 53, 35, 25, 17, 13$.

Example: In stage 1, the image is filtered with $ksize = 85$. Faces of Alice and Bob are annotated by workers. In stage 2, Alice and Bob are filtered with $ksize = 113$ while the rest of the image is filtered with $ksize = 53$. The face of Charlie is annotated by workers. In stage 3, Alice and Bob are filtered with $ksize = 113$, Charlie is filtered with $ksize = 85$, and the rest is filtered with $ksize = 35$. This proceeds accordingly with the values given above.

To ensure that the privacy of the people in the image is not compromised, the IntoFocus method uses the following methods to reduce the chances of inadvertent disclosure by inattentive workers or intentional disclosure by collusive workers.

Attention Check

To make sure that workers are performing the task correctly, the system contains a mechanism where the system knows

the locations of the faces in specific images in each set. If the worker did not detect all of the clearly visible faces, then the system would flag that worker and replace their work with another worker. This process is based on filtering outputs using ground truth comparisons (Marcus et al. 2012; Huang and Fu 2013). The reasoning behind the mechanism is to identify any and all workers that are not performing the task correctly. The attention check images were displayed in random order, so the workers would not know which images were evaluating the accuracy of their work. Due to the potential impacts associated with failure to prevent privacy disclosures, the system applies these mechanisms to uphold the promise of preserving the person's privacy.

Collusion Prevention

In related work, Kaur et al. (Kaur et al. 2017) demonstrated a system that segments an image and asks workers to choose regions that contain sensitive information. It is effective at reducing the risk of disclosure to individuals, but is vulnerable to coordinated group attacks. To mitigate this issue, our system uses anti-collusion protection to prevent workers from working together and redacting the entire picture together. The first protection is that no worker is able to work on the same image more than once, such as seeing the same image at different stages. The second protection that helps in preventing such an attack is that workers see images with all the previous redactions filtered out. Finally, to ensure that no one worker purposefully ignores a face, each image is presented to 3 workers and all their redactions are combined and filtered out. If one worker did not add an ellipse on a specific face, another worker will have the chance to do so, which in turn increases the privacy of the people in the image. For workers to be able to work together to extract all the information, they would need at least 21 different workers, and all 21 would need to accept the same task that contains the image they are trying to release unredacted.

Evaluation of IntoFocus

Our evaluation of IntoFocus tested the hypothesis that IntoFocus can acquire face redactions with 98% detection and 2% identification. Those levels affect the calculation of filter levels, but could be changed by users with greater or lesser tolerance for error. The experiment (Figure 5) was conducted on Mechanical Turk.

Treatment Condition (IntoFocus)

We used seven (7) stages (k sizes) for the IntoFocus method. Each stage was presented to at least 3 workers. Workers who previously worked on an image were excluded from seeing that image again. At the beginning of each stage, faces marked in previous stages were redacted to prevent identification by subsequent workers.

Dataset

The images being presented to the workers are from the IMDB-WIKI dataset (Rothe, Timofte, and Gool 2016) and the dataset that we collected to ensure the system is tested on real-life scenarios.

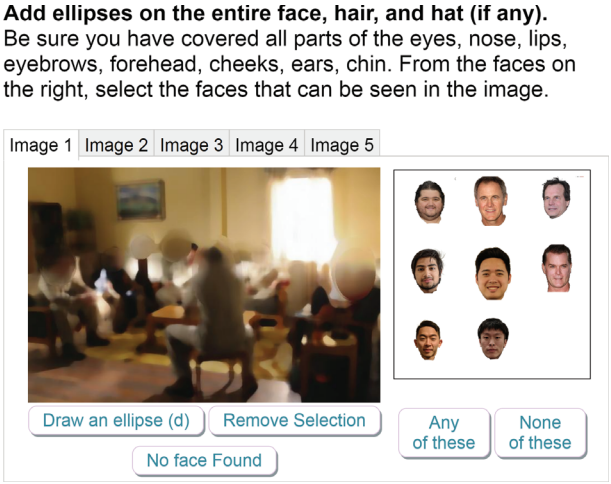

Figure 5: The IntoFocus task interface displaying an image in the process of redaction. A subtle difference in the filter level can be seen covering the faces of the people in the image. Here, the workers were required to perform two tasks on five different images. First, they add ellipses on all of the faces in the image. Next, they attempt to select the correct face that matches a person in the image. If they cannot perform the detection task, they were required to click the *No face Found* button. If they cannot identify the person in the image, they were required to select either the *Any of these* button or the *None of these* button.

A set of 186 different people were chosen to be identified by the workers. The tested images contained at least one person from the selected set. Recognition was quantified using k-anonymity (Sweeney 2002) with $k = 8$ among faces having similar non-facial attributes (skin tone, hair length, and hair color).

Image Presentation

There were 10 HITs, each containing five different images. Out of the five images, four were used for evaluation; one was used as an attention check. The images selected were different than the ones used in the perception study. The attention check image was not used as part of the evaluation process because it was chosen to fail the identity preservation test. The attention check images were randomly ordered among the treatment and control conditions. There were 10 image sets, 21 workers each, for a minimum of 210 assignments.

Task

Workers performed two tasks on each image. First, they annotated each face they could detect with a bounding ellipses. Second, they indicated if they could identify the face by selecting from among a panel of eight reference faces shown at the side. They could also indicate if they were unable to detect any faces and/or identify any of the detected faces.

Evaluation

The evaluation for the method and the system was performed using Mechanical Turk. A total of 10 different HITs were posted, covering 50 images (10 attention check images + 40 test images). The IntoFocus method was presented in 210 assignments. The system will be evaluated based on the ability to maximize detection and minimize identification.

Results

Out of 232 faces in the 40 images, 229 (98.7%) faces were detected. Microsoft Azure's face detection system detected 203 (87.5%).

Identification was measured with respect to only one face per image. Of the 840 trials (= 40 images $\times$ 21 workers), the correct identity was chosen only 7 (0.83%) times (7 distinct workers on 7 distinct images). This is much less than 105 (12.5%), as would be predicted by random chance (guessing 8 choices for each of 840 trials). Workers were instructed to select multiple faces if they could narrow the possibilities to a plural subset of the 8 choices, or click *Any of these* if they could not eliminate any of the choices. However, it is still possible that one or more of the correct selections were due to guessing.

The goal of the IntoFocus system is to improve automated systems. Based on that, it is the assumption that when crowd workers perform the face detection task, an automated system would be applied to detect the remaining faces. Coincidentally, the faces not detected by the IntoFocus system were all detected by Microsoft Azure's face detector. While Azure excels when the entire face is present, IntoFocus excels when the body is present. The faces that were not detected were all faces where the bodies were hidden from view and only the face was visible. On the other hand, IntoFocus detected all the faces that Azure failed to detect.

These results support our hypothesis that IntoFocus can acquire face redactions while limiting disclosure of *facial identities* to the workers, within the tolerance levels—detection (98%) and identification (2%)—assumed by our calculations of the median filter thresholds based on our perception study data.

Discussion and Future Work

The perception study of median filtered faces showed that people are truly different when it comes to detecting as well as identifying faces. Some participants were able to accurately pinpoint the location of faces when their peers were not able to detect anything until several filter levels later.

Machine-automated face detection remains a formidable problem. Even the best known algorithms achieve rates of only 78.8% (Zafeiriou, Zhang, and Zhang 2015) to 88.1% (Zhang et al. 2020). This is unacceptable for many privacy applications, in which there can be human consequences for any failure.

The actor dataset (Rothe, Timofte, and Gool 2016) used in the experiment was perfect for the study. To design a system that preserves the privacy of people, we need to understand how knowing the person in the image affects the worker's ability to identify them. The dataset gave us a chance to present workers with people they are familiar with. It also gave us images that had people in different sizes and in different regions of the image.

The IntoFocus method was not compared with other crowd-based methods, such as CrowdMask (Kaur et al. 2017), because they were generalized for multiple types of data. A comparative evaluation would be biased towards IntoFocus because it targets the specific application of face detection with protection of facial identities.

Future Work

The addition of machine learning and computer vision techniques would reduce the load on the workers as well as increase the productivity of the method. These techniques have the ability to automatically redact the easily visible faces and allow the people to focus on the occluded faces. Another direction would be to use a machine learning algorithm to assign different filter levels to different regions of the image. The image would then be sent out to crowd workers and they would be tasked with the redaction. This process would reduce the number of stages used by the IntoFocus method and significantly reduce the cost.

The current system focuses on faces because of the lack of a universal filter that withstands different types of image data (faces, texts, medical records, credit cards, etc.). A solution is an image filter that works with both text and faces, one that would hide the small details that help people with detection and still allows them to find the location of the text.

The IntoFocus system was lacking in some aspects. The system needed a stricter check for false positives—the current method relies on the attention check image. A possible solution is to run a face detector on each image and comparing the sizes of the two methods for the faces found by the face detector. Another issue is the image size, the method reduces the size of all the images before the redaction process. This causes smaller faces to not be redacted. Extra stages with lower filter levels can be added to redact these faces and the final redactions can be applied to the full sized image after the redaction process is complete, allowing the image to maintain the same starting image size.

Racial information from the workers that participated in the perception study were not collected. This information would show whether the selected filter levels accounted for any possible racial bias in the identification task. Racial information for the workers that participated in the system will be collected and applied in future iterations of the system.

Conclusion

This paper presented IntoFocus, a method that, given an image, uses crowdsourcing to redact the facial information that allows a person to be identified. IntoFocus adds to the knowledge of accurately redacting images while minimizing identity revelation to the worker and maintaining consistent results for different-sized faces. IntoFocus provides a method for other crowdsourcing applications (Bigham et al. 2010; Sorokin et al. 2010; Noronha et al. 2011) to preserve the privacy of facial information without affecting the functionality of the original application.

Acknowledgements

We are grateful to the anonymous crowd workers and lab participants, who made this research possible; to Dr. Chong Gu of the Statistical Consulting Service at Purdue University for his guidance and input on the analysis of results; and to Fanan Aleid, Bader Albassam, Nader Alawadi, Gaoping Huang, Meng-Han Wu, and Venkata Krishna Chaithanya Manam for their continuous support and guidance throughout the project.

This work was supported by a Google Faculty Research Award, Kuwait University and Sandia National Laboratories.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND2020-8742 C.

References

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; and Yeh, T. 2010. VizWiz: Nearly real-time answers to visual questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, 333–342. ACM.

Buolamwini, J., and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, 77–91.

Das, A.; Agrawal, H.; Zitnick, L.; Parikh, D.; and Batra, D. 2017. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? *Computer Vision and Image Understanding* 163:90–100.

Deng, J.; Krause, J.; and Fei-Fei, L. 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

Gross, R.; Sweeney, L.; Torre, F. d. l.; and Baker, S. 2006. Model-Based Face De-Identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 161–161.

Gurari, D.; Jain, S. D.; Betke, M.; and Grauman, K. 2016. Pull the Plug? Predicting If Computers or Humans Should Segment Images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, 382–391.

Hardie, R. C.; Barnard, K. J.; and Armstrong, E. E. 1997. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing* 6(12):1621–1633.

Huang, S.-W., and Fu, W.-T. 2013. Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, 639–648. San Antonio, Texas, USA: Association for Computing Machinery.

Hummel, R. A.; Kimia, B.; and Zucker, S. W. 1987. Deblurring Gaussian blur. *Computer Vision, Graphics, and Image Processing* 38(1):66–80.

Jain, V., and Learned-Miller, E. 2010. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst.

Kajino, H.; Baba, Y.; and Kashima, H. 2014. Instance-Privacy Preserving Crowdsourcing. In *Second AAAI Conference on Human Computation and Crowdsourcing*, HCOMP '14.

Karnin, E. D.; Walach, E.; and Drory, T. 2010. Crowdsourcing in the Document Processing Practice. In *Current Trends in Web Engineering*, Lecture Notes in Computer Science, 408–411. Springer, Berlin, Heidelberg.

Kaur, H.; Gordon, M.; Yang, Y.; Bigham, J. P.; Teevan, J.; Kamar, E.; and Lasecki, W. S. 2017. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, 1301–1318. New York, NY, USA: ACM.

Kolmogorov, A. N. 1956. *Foundations of the Theory of Probability: Translation Edited by Nathan Morrison. With an Added Bibliography by AT Bharu-cha-reid*. Chelsea Publishing Company.

Lasecki, W. S.; Song, Y. C.; Kautz, H.; and Bigham, J. P. 2013. Real-time Crowd Labeling for Deployable Activity Recognition. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, 1203–1212. New York, NY, USA: ACM.

Lasecki, W. S.; Gordon, M.; Teevan, J.; Kamar, E.; and Bigham, J. P. 2015a. Preserving Privacy in Crowd-Powered Systems. In *In AAMAS 2015 Workshop on Human-Agent Interaction Design and Models*, AAMAS '15.

Lasecki, W. S.; Gordon, M.; Leung, W.; Lim, E.; Bigham, J. P.; and Dow, S. P. 2015b. Exploring Privacy and Accuracy Trade-Offs in Crowdsourced Behavioral Video Coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 1945–1954. New York, NY, USA: ACM.

Lasecki, W. S.; Teevan, J.; and Kamar, E. 2014. Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, 248–256. ACM.

Lewis, M. B., and Edmonds, A. J. 2003. Face Detection: Mapping Human Performance. *Perception* 32(8):903–920.

Lindsay, D. S.; Jack, P. C.; and Christian, M. A. 1991. Other-

race face perception. *The Journal of Applied Psychology* 76(4):587–589.

Marcus, A.; Karger, D.; Madden, S.; Miller, R.; and Oh, S. 2012. Counting with the crowd. *Proceedings of the VLDB Endowment* 6(2):109–120.

Nada, H.; Sindagi, V. A.; Zhang, H.; and Patel, V. M. 2018. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. *arXiv preprint arXiv:1804.10275*.

Noronha, J.; Hysen, E.; Zhang, H.; and Gajos, K. Z. 2011. Platemate: Crowdsourcing Nutritional Analysis from Food Photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, 1–12. New York, NY, USA: ACM.

Quinn, A. J., and Bederson, B. B. 2011. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, 1403–1412. New York, NY, USA: ACM.

Rothe, R.; Timofte, R.; and Gool, L. V. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*.

Sorokin, A.; Berenson, D.; Srinivasa, S. S.; and Hebert, M. 2010. People helping robots helping people: Crowdsourcing for grasping novel objects. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS '10, 2117–2122.

Swaminathan, S.; Fok, R.; Chen, F.; Huang, T.-H. K.; Lin, I.; Jadvani, R.; Lasecki, W. S.; and Bigham, J. P. 2017. 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*, UIST '17, 807–815. New York, NY, USA: ACM.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570.

Varshney, L. R. 2012. Privacy and Reliability in Crowdsourcing Service Delivery. In *2012 Annual SRII Global Conference*, 55–60.

von Ahn, L.; Liu, R.; and Blum, M. 2006. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, 55–64. New York, NY, USA: ACM.

Zafeiriou, S.; Zhang, C.; and Zhang, Z. 2015. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding* 138:1–24.

Zhang, B.; Li, J.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Xia, Y.; Pei, W.; and Ji, R. 2020. ASFD: Automatic and Scalable Face Detector. *arXiv:2003.11228 [cs]*. arXiv: 2003.11228.

Zhu, C.; Zheng, Y.; Luu, K.; and Savvides, M. 2017. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep Learning for Biometrics*. Springer. 57–79.