# Automated Bystander Detection and Anonymization in Mobile Photography

David Darling[1], Ang Li[2(✉)], and Qinghua Li[1]

[1] University of Arkansas, Fayetteville, USA
{dwdarlin,qinghual}@uark.edu
[2] Duke University, Durham, USA
ang.li630@duke.edu

**Abstract.** As smartphones have become more popular in recent years, integrated cameras have seen a rise in use. This trend has negative implications for the privacy of the individual in public places. Those who are captured inadvertently in others' pictures often have no knowledge of being included in a photograph nor have any control over how the photos of them might be distributed. To address this growing issue, we propose a novel system for protecting the privacy of bystanders captured in public photos. A fully automated approach to accurately distinguish the intended subjects of photos from strangers is first explored. To accurately distinguish these subjects and bystanders, we develop a feature-based classification approach utilizing entire photos. Additionally, we consider the privacy-minded case of only utilizing local face images with no contextual information from the original image by developing a convolutional neural network-based classifier. Considering the face to be the most sensitive and identifiable portion of a bystander, both classifiers are utilized to form an estimation of facial feature locations which can then be obfuscated to protect bystander privacy. We implement and compare three methods of facial anonymization: black boxing, Gaussian blurring, and pose-tolerant face swapping. To validate and explore the viability of these anonymization methods, a comprehensive user survey is conducted to understand the difference in appeal and viability between them.

**Keywords:** Privacy · Mobile photography · Facial anonymity · Face swapping · Obfuscation

## 1 Introduction

Digital photography is an enormously growing trend among the general public spurred primarily by the prevalence of smartphones with cameras in the daily lives of users. Attempts at estimating the number of digital photos captured annually have shown significant increases year over year with no sign of slowing down. In 2016, 660 billion photos were estimated to have been captured which

Ang Li was a PhD student at the University of Arkansas at the time of this work.


N. Park et al. (Eds.): SecureComm 2020, LNICST 335, pp. 402–424, 2020.
https://doi.org/10.1007/978-3-030-63086-7_22

increased to 1.2 trillion by 2017. Additionally, smartphones were estimated to be primarily responsible for these photos, as they captured roughly 85% of the photos taken. Digital cameras, the next largest device group, made up only 10.3% by comparison [23]. This significant increase in digital photography demonstrates that people are more likely to be taking photos across many scenarios, including in public locations where oftentimes other people are nearby. Due to the crowded nature of public locations, strangers are frequently included in personal photos by circumstance. Figure 1 provides an example of such an image taken in public including both a target person and strangers. These strangers, or bystanders, often are completely unaware of photos being taken of them. Even realizing that a photo was taken, individuals usually have very little recourse to ask to be removed from the photo or have the photo deleted in a public setting.

Extending a preliminary workshop version [7] in which we proposed feature sets and used them to detect bystanders and targets, this work explores beyond feature-based solutions for bystander detection and studies effective ways to obfuscate bystander faces. To this end, we propose both a feature-based classification approach utilizing entire photos and a privacy-minded, convolutional neural network (CNN)-based approach utilizing only local face images with no contextual information from the original photo to investigate how different models learn the intrinsic visual differences between targets and bystanders. In order to train and evaluate our models, a real-world dataset consisting of over 200 photos and over 500 faces has been created to provide a generalized representation of the types of images that commonly can be found uploaded to social media. Generally, the photos provide a mixture of both celebrities appearing in public with bystanders behind them and typical people in front of landmarks or other locations of interest with strangers also inadvertently captured in the photo.

In addition to the classifiers, methods for effective facial obfuscation also are implemented and examined through a user study (with IRB approval) as a part of this work. The methods for anonymizing faces not only include standard methods such as Gaussian blurring and black boxing but also a novel approach of face swapping using a state-of-the-art position map regression network [8]. To gain a more full understanding of how these obfuscation methods impact and are perceived by end users, an in-depth survey is carried out and evaluated to determine the preferred methods of actual users. General opinions on privacy as it relates to digital photos capturing strangers are also collected to validate the assumptions behind this work.

This paper's main contributions are summarized as follows:

- We propose an automated system for protecting bystander's privacy in mobile photography with a unique feature that it can work as a standalone tool on a user's smartphone, without relying on inter-user interaction or any online platform which are commonly needed by previous solutions.
- We propose a novel feature-based classification approach utilizing entire photos and a privacy-minded CNN-based approach utilizing local face images for automatically distinguishing targets from bystanders in mobile photography.

**Fig. 1.** Example image with target clearly featured in the foreground and bystanders included to the left and right in the background. Image source: http://mensstreetfashion.weebly.com/home/kanye-west-style

The two approaches are evaluated and compared to explore the tradeoff between the distinguishing accuracy and privacy.
– We implement three face obfuscation methods for complete face anonymization including black boxing, blurring, and face swapping, and carry out a user study with 89 respondents to evaluate opinions on bystander privacy and on the acceptability of the three face anonymization methods.

The remainder of this paper is organized as follows. Section 2 provides an overview of related works. Section 3 presents an overview of the design for the system. Section 4 describes the feature-based approach for target/bystander classification. Section 5 describes the convolutional neural network approach. Section 6 provides evaluations of both models. Section 7 provides a technical description of proposed facial obfuscation methods and presents the results from the user survey. Section 8 concludes this paper and discusses future work. Last section offers acknowledgments.

## 2 Related Work

Works related to improving facial privacy in photography have previously followed trends such as utilizing photographer-bystander cooperation. Li et al. [17,18] design systems for cooperation between smartphone users to blur requesting users' faces. Jung and Philipose [13] utilize a method of gesture recognition to detect a person who wants to be excluded from a photo. *MarkIt* [21] can

perform automated covering of user-defined objects in photos, but users must manually predefine objects to be hidden. This work, by comparison, explores a fully automated approach towards the identification of bystanders to be obfuscated and requires no manual interaction from photographers or those captured in photos.

Other approaches in automated anonymization require users interested in having their privacy protected to wear specialized markers. Schiff et al. [24] propose a scheme whereby specialized markers that users wear can be recognized by an automated system which will then blur their faces. Bo et al. [4] similarly utilize worn QR-codes on clothing to automatically determine individual privacy preferences. Our approach does not require any worn markings; only visual attributes of persons within individual images are used to anonymize those who are inadvertently captured.

Some works explore photo privacy protection for individuals in online social network settings [12,19,26] or in individual phones [16]. Li et al. proposed *HideMe* [19], a system for defining scenarios for photo access control and distance-based face blurring. Xu et al. [26] developed a facial identification system to incorporate captured persons into the decision process of sharing photos. Ilia et al. [12] proposed fine grained access control based on face detection to prevent individual faces from being viewed by other users. Each of these works depend upon either accurate facial recognition to identify bystanders for anonymizing or specific scenario definitions from the photographer. Our work circumvents the need for both face recognition and input from the photographer by automating the detection of bystanders and providing low-impact solutions for obfuscating their faces.

One recent work in parallel to ours also explores detection of bystanders utilizing features computed over individuals [10]. However, the features used in our work are different from theirs. Also, that work focuses entirely on feature engineering and effectiveness of predictive models, but our work presents a full system for both detecting and anonymizing bystanders and explores the trade-off between bystander protection and photo quality/usability. Additionally, we consider privacy-aware machine learning with the development of a CNN which does not require access to an entire user photo and can predict solely based on facial images. While Hasan et al. [10] explore both transfer learning on deep neural networks, our approach utilizes a simplified feature-set which allows for simpler and smaller classifiers.

## 3 System Overview

Our proposed system works on the photographing phone and consists of four main processes that can automatically run over taken photos. The first is an initial pass of face detection which identifies face regions and positions. This process of facial detection is fully automated and requires no manual user input specifying specific photos or regions to be focused by the algorithm. Facial region data from this pass can then be forwarded either to our feature-based or CNN classifiers which automate the detection of targets and bystanders. The classifications

along with face landmarks are then used to perform obfuscation processing with the black boxing, blurring, and face swapping methods which can be selected by the user as a system configuration parameter. The resulting anonymized photo is the final output. Figure 2 provides a visualization of this system. We find that both classifiers are robust against nonstandard scenarios where a photo might not include any bystanders or any human targets (pictures of scenery are an example of this.)

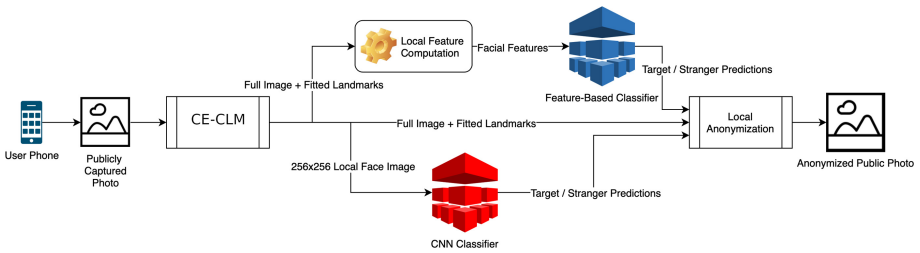


**Fig. 2.** Architecture of the system.

## 4 Feature-Based Bystander Classifier

### 4.1 Feature Identification

In order to begin designing a model for classification of an abstract concept such as who is the desired target of a photo, it is essential to correctly break the problem down into quantitative measurements which provide a suitable amount of information to distinguish classes. Relative face size (the size of a given face relative to the maximum face size in the image) and face deviation from center are identified to model the fact that bystanders are often included in the background or periphery of photos. Local blurriness of a face is also found to regularly indicate a person was not the intended focus of a picture. Beyond these metrics of relative visual differences, additional features are identified to model the fact that bystanders are usually unaware of having their photo captured. Additionally, non-relative features are needed to account for scenarios where photos might exclude any human targets (for example bystanders are captured in an image of scenery or some other non-human target). Head pose angles and gaze deviation from the camera are decided as good indicators of this, as bystanders will commonly not be looking at the camera capturing them if they are unaware of it. Considering all of these metrics in a single model, relative face size, face deviation from photo center, local blurring, head pose, and gaze deviation are identified as a sufficient number of features to capture the complexity of the target/bystander classification problem.

### 4.2 Feature Extraction and Computation

**Relative Face Size.** In order to accurately capture face size in an image, traditional bounding box methods used by most state-of-the-art object detection models such as YOLOv3 [22] are not sufficient. These bounding boxes normally have no guarantee of forming a tight bound on the object in question. Instead, we recognize that recent advances in constrained local neural fields (CLNFs) for facial landmark detection are more suited for the task of robust landmark placement for photos taken in the wild due to their resistance to factors such as pose differences, lighting changes, and local facial differences such as hair or accessories. To this end, we adopt a Convolutional Experts Constrained Local Model (CE-CLM) [1,27] to perform accurate facial landmark placement. These types of models function by first capturing landmark shape variations with a point distibution model (PDM) and then modeling the local differences in the visual appearances of fitted landmarks with the use of local patch experts. This model, pretrained over LFPW [3] and Helen [15] training sets, is able to accurately determine landmark positions and provided higher detection rates on smaller faces as well as partially occluded faces as compared to more popular systems such as Dlib [14] during our experimentation.

Figure 3 provides a visualization of the standard collection of facial landmarks which are fitted by networks such as Dlib and the CE-CLM model. Refer to this figure for locating any numbered facial landmarks mentioned subsequently. Once a tight bound has been formed around each face using these landmarks, the maximum size of all faces found in the photo is calculated as:

$$\max_{0 \leq i \leq n} S(i) = (\alpha_i - \beta_i)(\gamma_i - \delta_i) \tag{1}$$

where $i$ refers to the index of the current face in the photo which can have $n$ faces, $\alpha_i$ is the x-coordinate of facial landmark 16, $\beta_i$ is the x-coordinate of landmark 0, $\gamma_i$ is the y-coordinate of landmark 24, and $\delta_i$ is the y-coordinate of landmark 8. The subscript $i$ for each of these variables indicates that these are specific to the face with index $i$. Similarly, the relative face size metric for each face is then calculated simply as:

$$R(x) = \frac{s_i}{\mu} \tag{2}$$

where $s_i$ is the size of face $i$ and $\mu$ is the maximum face size computed in Eq. 1 Thus, the relative size of any given face will always fall within the bounds of $(0, 1]$.

**Deviation of Face from Center.** Face position within an image can be extracted and computed in a method similar to the relative face size metric. Facial landmarks are fitted to each detected face in an image with the center of the face treated as landmark 33. By utilizing this landmark, it is possible to accurately extract what region of the image a given face is located assuming the dimensions of the image are known. To create a useful metric for supervised

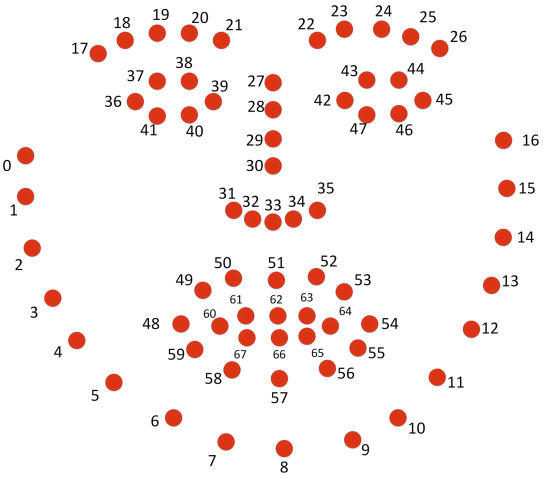


**Fig. 3.** Standard template of facial landmarks fitted by common face detection networks.

learning models, the face position is computed as the amount of deviation from center of the image. The intuition here is that the feature generally should (but not always) provide a positive correlation with the likelihood that a given face belongs to that of a stranger. To better capture the meaning behind a person's face deviation from the center of a photo, the metric should be normalized such that its range of values carry the same meaning across different photo sizes. To this end, the deviation of a detected face is calculated as:

$$D(i) = \frac{|\epsilon_i - \zeta|}{w} + \frac{|\eta_i - \theta|}{h} \tag{3}$$

Here, $\epsilon_i$ is the x coordinate of landmark 30, $\zeta$ is the x-coordinate of the computed center of the image, $\eta_i$ is the y-coordinate of landmark 30, $\theta$ is the y-coordinate of the center of the image, $w$ is the width dimension of the image, and $h$ is the height dimension of the image. The subscript $i$, as in previous equations, refers to the $i$-th detected face in the image. The deviation in both x and y coordinate axes is summed for this metric because their relative importance towards determining the likelihood of a person being a bystander in an image is unclear. For example, consider a case where a target is featured centrally in front of public stairs, and a bystander is captured farther up the stairs in the image. Although the bystander might be centrally located in the x-axis, their y-axis deviation is more important in this case. Because of this ambiguity in importance across image cases, both axis deviations are weighted the same when computing the overall deviation. The bound of this metric then becomes $[0, 2]$.

**Local Blurring.** Many methods exist to compute the amount of blurring that occurs over a localized region in an image. One of the most popular methods is

to compute a Fast Fourier Transform over an image to break it down into its constituent frequencies and perform frequency domain analysis on the results. This method is not ideal for generating a general metric of blurriness across photos as it is difficult to identify the specific frequencies in the general case which mark a region as blurry vs. another. Instead, the convolution of the Laplacian method proposed by Pech-Pacheco et al. [20] is selected for its ability to provide average edge variance in an image as a single floating point result. This is desirable from a feature engineering perspective, because blurriness, as a measure of edge variance, effectively captures the desired information from an image region without requiring a secondary classifier to convert frequency component information into a blurriness boolean.

To perform this method, the Laplace operator which is defined in 2-dimensional Cartesian space canonically as:

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{4}$$

However, in a discrete grid such as an image, the discrete Laplacian is used which essentially is a convolution of the following kernel:

$$L = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{5}$$

This operation is therefore used to generate a floating point "rating" of the blurriness over each local region of detected faces defined by their bounding boxes extracted by padding the space around the outer facial landmarks. Bounding boxes are used in this case rather than the tight bounds featured in previous features because the wider region around the face contains additional edges for the Laplacian convolution operation. This provides a better idea of how much variance is in the face's surrounding edges.

**Head Pose Estimation.** Head pose defines the way an individual's head is oriented in 3-dimensional space. Intuitively, being able to extract some measurements of head orientation should capture the trend of strangers having their heads turned away from a capturing camera. To define head orientation, 3 main parameters are needed:

- Pitch: This defines the angle a head is looking up or down. Essentially a measure of vertical tilt.
- Roll: The angle a head is tilted from side to side. Note that this is distinct from yaw in that roll can vary while the face remains looking forward.
- Yaw: The angle a head is turned from left to right. For example, when turning to look at something behind a person, the head yaw angle becomes more extreme.

Each of these angles are intrinsically within the CE-CLM models as they internally keep a 3D representation of the fitted face landmarks. These 3D representations can be utilized to estimate accurate head pose information by solving the n-point in perspective problem [1,2,27]. This procedure essentially allows for accurate estimation of head pose angles in general images. These angles are represented as continuous floating point values which can range both in positive and negative directions.

**Gaze Deviation.** This feature is intended to provide additional information about whether a given person has awareness of being included in a photograph. Although head pose angles can provide information whether the face in question is oriented toward the camera, these angles do not tell the whole story. A stranger in a photo could very well have their head oriented toward the camera if, for example, they are walking behind someone taking a selfie. To provide more information to learning models about these situations, gaze deviation from the center of the camera focus could be included.

In order to extract gaze angles, a slightly altered method of utilizing CLNFs proposed by [25] is adopted. In the original work, CLNFs were utilized to form a PDM of the eye landmarks for synthesized eyes. This estimation of the shape of the eye can then be used to further estimate what direction the eye is oriented in (and where the eye is looking). The accuracy of the gaze estimation was improved utilizing pre-determined intrinsic camera parameters. However, because photos captured by mobile devices in the wild provide no information about these parameters, we adopt the model to utilize some default values for these camera parameters. These include estimated distance of the face from the camera, focal length, and optical center. This sort of calibration cannot be done for images collected for the training dataset as there simply are too many camera-to-person-positionings to take into account. Therefore, the gaze vectors should be considered rough estimates of whether a given person is looking in the general direction of a camera. The rationale behind this is that a binary value of looking vs. not looking at a camera is still a desirable trait in identifying bystanders in a mobile photo. Example gaze vectors plotted over an image can be seen in Fig. 4.

By utilizing extracted gaze vectors, it is possible to trace a ray from the center of a given face into the estimated camera location. Then, by organizing world coordinates with the camera at the origin $(0, 0, 0)$, it is possible to estimate the position of a face from the camera using coordinates. The tracing process begins by setting up the following equation:

$$\kappa_z + u\boldsymbol{\lambda}_z = 0 \tag{6}$$

Here $\kappa_z$ refers to the z-coordinate or estimated depth of facial landmark 30, $u$ is a scaling factor which must satisfy the equation, and $\boldsymbol{\lambda}_z$ is the z-coordinate vector of the average gaze angle for a given face. Solving this equation for scaling factor $u$ allows us to then scale the other components of the full gaze vector $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda}' = u\boldsymbol{\lambda} = u \begin{bmatrix} \boldsymbol{\lambda}_x \\ \boldsymbol{\lambda}_y \\ \boldsymbol{\lambda}_z \end{bmatrix} \tag{7}$$

To compute the point $\rho$ where the ray traced from the face intercepts with the origin or camera z-plane, the scaled gaze vector must be added to $\kappa$ (landmark 30):

$$\rho = \kappa + \boldsymbol{\lambda}' \tag{8}$$

The value of $\rho$ enables a deviation estimate for anyone looking in the direction of the camera's z-plane. Using the x and y components of $\rho$, and assuming the camera is the origin of the world coordinates, we can use a simple 2-D distance formula calculation to find the final deviation:

$$D = \sqrt{\rho_x^2 + \rho_y^2} \tag{9}$$

The resulting value of $D$ computed for any given detected face is the final metric for a supervised learning model.

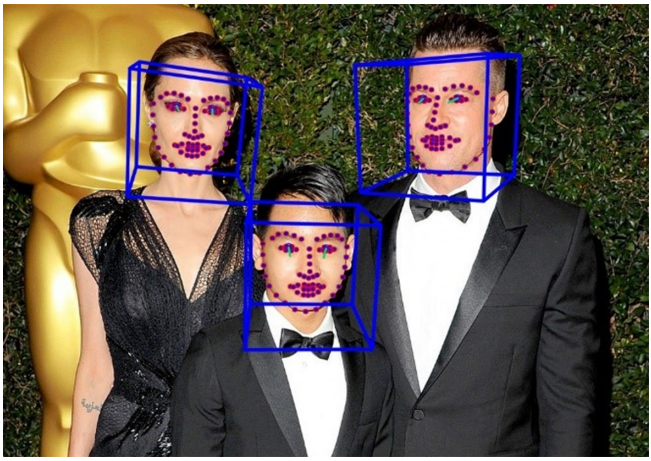

**Fig. 4.** Gaze vectors from the CLNF model in green with face landmarks from the CE-CLM model on an example image with no bystanders. Original image by Steve Granitz, WireImage. (Color figure online)

### 4.3 Supervised Learning Model Consideration

Because all of the facial features are collectively designed to fully capture the complexity of the bystander classification problem, they should generalize very well to a wide range of supervised learning models. In our implementation, a collection of diverse classifiers are implemented and evaluated, including Gradient Boosted Decision Tree, Multilayer Perceptron, Random Forest, and Support Vector Machine. Detailed descriptions of these algorithms as well as discussions of their effectiveness in learning from the computed features are presented in the evaluation section.

## 5 CNN-Based Bystander Classifier

The feature-based bystander classification method needs the entire photo to work, since it considers a target's or a bystander's face in the context of the photo (e.g., whether a face is in the center region of the photo or not). An interesting question is whether it is possible to distinguish target and bystander just based on their faces without relying on the context in the photo. Basing on face only could result in a more privacy-minded classifier which would only require images of faces without revealing their surroundings or even other people nearby. Such a classifier would be especially useful for any application where users might utilize a cloud-based solution for automated bystander training and detection, since the bystander's face itself does not leak his/her privacy (e.g., location) when uploaded to the cloud for analysis. To this end, a CNN architecture is designed and validated on the same dataset as the feature-based models.

### 5.1 Network Architecture

Figure 5 provides an overview of the complete network architecture. The network utilizes increasingly small convolutions separated by max pooling layers. The activation functions utilized by the network are rectified linear unit (Relu) for the two convolutional layers and sigmoid for the final activation layer. The dense layers in the latter portion of the model are intended to produce meaning from the large feature vector and condense them into more usable, countable features. The final dropout layer is included to reduce overfitting on the training set. It is set to drop inputs at a rate of 0.25 which experimentally achieved best results. The filter size of the convolutional layers is small, at a size of (2,2). This is to ensure that fine-detailed features such as eye direction might be captured, as eyes in the facial dataset can sometimes be very small (only being formed from a few 10s of pixels). Stride for each of these kernels is set to (1,1), such that a direct sweep of the kernel is performed over the image.

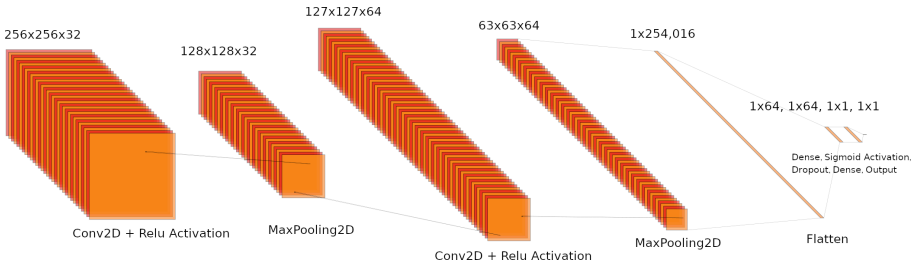


**Fig. 5.** Overall architecture of the convolutional neural network.

## 6 Model Evaluation

### 6.1 Dataset

The dataset for this problem was created specifically for this project due to the fact that no existing datasets could be found at the time which provided the types of required images. The images in the set were automatically collected from freely available online sources including social media platforms, public news sites, and image repository sites. Collected photos were manually reviewed to ensure the dataset would be able to provide a good generalization of the types of images that might be encountered in the wild. To be included, photos were required to have at least one human face. Photos could have all targets, all bystanders, or a mixture of both. The photos are all taken in various settings including indoor/outdoor locations, public venues with varying degrees of crowding, and daytime/nighttime lighting conditions. Photos excluding human targets such as scenery photos were also included to ensure the developed models could generalize to these challenging scenarios. In total, the dataset consists of 515 valid face images extracted from 222 photos. It is worth noting that our dataset is of comparable size to Hassan et al. [10] (515 facial images vs 600). In order to allow other researchers to utilize these images and contribute new images, we have made the dataset publicly available at [6].

### 6.2 Feature-Based Bystander Classification

**Model Selection and Implementation.** The following supervised learning algorithms were selected and implemented to provide a good coverage of the varieties of popular classifiers and demonstrate that even with multiple different algorithms, our features generalize well.

*Gradient Boosted Decision Tree (GBDT).* GBDTs are an enhancement over normal decision trees whereby an ensemble of weaker models are utilized to form a single classifier [9]. The number of estimators or trees used was 300. A maximum depth of 10 was also selected for the classifier. The learning rate was set to 0.03 after multiple training attempts.

*Random Forest (RF).* Instead of the weak models favored by gradient boosting methods, the RF approach is to make use of deep, fully grown trees and average them together to reduce variance and overfitting [5]. The RF algorithm was selected to offer a good comparison with GBDTs. The number of estimators for this algorithm was selected as 50 with a max depth of 6.

*Support Vector Machine (SVM).* Because SVMs are considered very suitable for binary classification [11], they were chosen for evaluation alongside other more advanced algorithms. Best results were achieved with a linear kernel, with hyperparameter C set to 10 and gamma set to 0.001.

*Multilayer Perceptron (MLP).* Neural networks are a logical choice for a feature-based model such as this. The chosen architecture for the MLP is 3 hidden layers of size (7,5,3). The hyperbolic tangent function was selected for activation, and a learning rate of 0.03 is used with alpha parameter set to 0.0001. After multiple training attempts, this coupling of architecture and hyperparameter values provided best results.

**Training and Evaluation.** Each of the feature-based algorithms was trained over a random 80/20 train/test sample split of 515 feature sets (one set for each face image in the dataset). Validation metrics are shown in Table 1. Of the classifiers we trained, the GDBT and MLP neural network were able to achieve the best validation accuracy at 94.34%. The RF and SVM were still able to achieve acceptable accuracy over 90%. The reasoning for this is that the simpler SVM model had a higher tendency to overfit the training set and resulted in less generalizable models. The RF model suffered from the lower depth of underlying decision trees relative to the GBDT, but required less processing to perform prediction passes.

Different classifiers were able to predict targets and bystanders with varying degrees of effectiveness. This is shown by the precision, recall, and F-1 score of each algorithm. Precision is defined as the number of true positives out of the combined number of true positives and false positives. Recall is defined as the number of true positives out of the combined number of true positives and false negatives. F1-score is the harmonic mean of the two. The MLP actually had the overall highest F1 scores with 0.93 (target positive) and 0.94 (bystander positive). The GDBT actually suffered from significantly worse precision when predicting targets which indicates that it might have difficulty identifying relevant targets in the wild. It actually acheived the same target F1 score as the SVM model which had the lowest scores for both targets and bystander prediction.

In order to determine how effective each of the engineered features was individually to the classifiers, the algorithms were trained over different subsets of the complete feature-set. Each subset had one feature removed and the others included. Table 2 lists the validation accuracies for these models. Of the features tested for exclusion from the models, the gaze deviation metric, when removed, had no real impact on the performance of the GBDT and only a small impact to the other models. By contrast, both the face size and center deviation metrics had significant harmful impacts on all of the classifiers. This could be an indication that the inherent inaccuracy of the gaze metric itself was a problem for training. Additionally, the gaze metric could simply be redundant where head pose information might have been sufficient. For these reasons the gaze deviation metric could be a candidate for removal in the interest of improving training and prediction speed. Appendix A provides a detailed overview of runtime performance for all classifiers. Because of the streamlined set of features used by the classifier, the computational complexity of all classifiers were kept low and should be suitable for direct implementation on resource constrained mobile devices.

**Table 1.** Accuracy, precision, and recall metrics for all classifier models. (T) and (B) indicate that the metric was computed with target or bystander respectively as the positive class.

| Learning Algorithm | Validation Accuracy | Precision (T) | Recall (T) | F1 Score (T) | Precision (B) | Recall (B) | F1 Score (B) |
|---|---|---|---|---|---|---|---|
| MLP | 94.34% | 0.98 | 0.88 | 0.93 | 0.90 | 0.98 | 0.94 |
| SVM | 90.57% | 0.94 | 0.87 | 0.90 | 0.87 | 0.94 | 0.90 |
| RF | 92.45% | 0.93 | 0.95 | 0.94 | 0.94 | 0.92 | 0.93 |
| GDBT | 94.34% | 0.83 | 1.00 | 0.90 | 1.00 | 0.88 | 0.94 |

**Table 2.** Validation accuracy for models trained using feature-subsets

| Learning Algorithm | No Gaze Deviation | No Face Size | No Center Deviation | No Head Pose |
|---|---|---|---|---|
| MLP | 90.57% | 77.36% | 81.13% | 85.71% |
| SVM | 92.45% | 81.13% | 84.91% | 86.68% |
| RF | 92.45% | 81.13% | 77.36% | 83.81% |
| GDBT | 94.34% | 84.91% | 83.02% | 84.76% |

### 6.3 CNN-Based Bystander Classification

In order to train the CNN, the same image set was used as for the feature-based model with a similar 80/20 train/test split. However, extracted local face images were used rather than entire photos. As mentioned previously, the goal of analyzing this network is to see if a privacy-concerned model could function well without needing to process entire images. Actual training took place utilizing mini-batch gradient descent with a batch size of 24 samples and 17 steps per epoch. Through experimentation, it was found that the CNN training accuracy generally converged after 15 training epochs. The testing accuracy was 81.55% with target precision and recall of 82.69% and 81.13% respectively which is significantly lower than the best feature-based models. Additionally, the predictive runtime of the model was found to be much higher than the feature based models at around 244ms on an Intel i9 platform (see Appendix A Table 3 for detailed runtime comparisons). However, these metrics are still impressive considering the loss of contextual information about a photo that the CNN experiences in comparison with the feature sets.

Originally, it appeared as though the networks would quickly converge during training due to the relatively stable loss value that was reported for the first 10 training epochs. However, by doubling the training epochs, it was found that true loss convergence did not generally happen until further training occurred. Through multiple experiments, the number of epochs that resulted in the lowest loss without overfitting and harming test accuracy occurred with 15 as mentioned previously. An example of a complete training sequence is provided in Fig. 6 with

loss plotted. Additionally, Fig. 8 in Appendix A plots training accuracy over the same number of mini-batches. It is interesting to note that just three epochs were enough to provide significant improvements in loss and training accuracy, but in our experiments the model was not able to effectively generalize to the test set after such short training periods. It is also worth noting that with a training accuracy approaching 89% and test accuracy reaching 81.55%, overfitting still occurs in the model even with a reasonable dropout rate of 0.25.

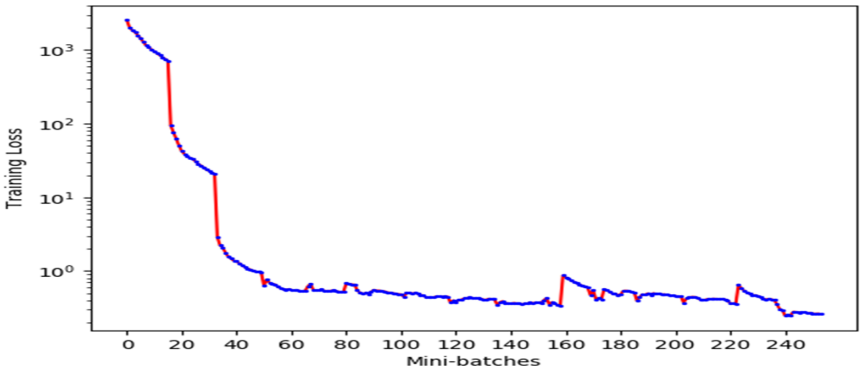


**Fig. 6.** Progressive loss of the network for each mini-batch.

The main contribution of this model is the fact that, unlike the feature-based approach which will require users to supply entire images to compute contextual features such as relative face size, this model is still able to effectively distinguish bystanders and targets with only face images, as it convolves over a "cropped" facial region. The hope is that users who might not want to supply entire photos to an automated system especially when the system is hosted in the cloud can still make use of a less invasive model which will not see any exposing information at the cost of a significant, flat decrease in performance in terms of accuracy, precision, and recall.

## 7 Anonymizing Bystander Faces

For a complete system to ensure the privacy of strangers captured in images, it is necessary to automate the obfuscation of faces so that the ease of identifying someone is severely limited. Although there are some methods available to obfuscate faces, it is still unclear how acceptable they are to users. In this section, we implement three different methods for facial anonymization and explore them with a user study. Figure 7 provides a visualization of each method on an example photo from the training dataset.

**Fig. 7.** From left to right: original image, image with black boxing, image with blurring, image with face swapping. Image source: https://cdn.ebaumsworld.com/2008/08/861915/phelps01.jpg

### 7.1 Implementation of Obfuscation Methods

*Black Boxing.* Black boxing of the face is the simplest, and arguably most secure method for facial anonymization. The detected stranger face is completely removed from the image with every pixel RGB value being set to black. In this way, there is not any remaining information which can be gleaned from the face, such as race or face size which the other proposed methods can leave behind. Although this provides the strongest guarantee of privacy, the visual impact to a photo can be very harmful depending on the surrounding lighting conditions.

*Gaussian Blurring.* Blurring is considered an intermediary between the black boxing and face swapping methods in terms of intended impact to photo quality. Blurring on smaller faces can be relatively unnoticeable in images, especially on persons captured in the far background of an image. Unlike black boxing, information such as race and even hair color can still be preserved depending on the resolution of the photo. However, facial features are always guaranteed to be completely anonymized. Gaussian blurring is used for this system due to its popularity and ease of computation. A kernel size of (70,70) is used and achieves acceptable blurring.

*Face Swapping.* Pose-tolerant face swapping traditionally required the use of deep CNNs such as the deepfake project which requires specific training for the two faces attempting to be swapped. However, recent advances in automated swapping, specifically the introduction of position map regression networks, have allowed for excellent generalized swapping of 3D face masks without any need for targeted training. Using an implementation of this method introduced by Feng et al. [8] allows for a novel technique of anonymizing bystander faces with any selection of "public" faces. These public faces could be commonly-known celebrities or even artificially generated portraits. Assuming the face is realistic enough and lighting differences are not too extreme, the results can be very believable. For this project, faces taken from stock photos found online were

used. To match skin tone for each detected stranger face, it is possible to compute an average pixel color value utilizing facial landmarks within the face region to gain a representation of their overall face color. This color is then compared with the precomputed averages of a collection of public faces. The public face which minimizes the difference is selected for swapping.

### 7.2 Survey of Users on Face Anonymization

In order to validate our obfuscation methods and demonstrate that they are both effective at anonymizing faces and cause limited impact to user photos, we conduct a comprehensive user survey.

**Questions.** To gain a better understanding of how actual users regard the protection of privacy for strangers and themselves in photos, one portion of the survey asked participants for their opinions on a series of questions relating to digital photo privacy and stranger protection. These questions were presented as:

- Question 1: Ensuring the privacy of digital photos is important.
- Question 2: You would want your privacy protected if someone took a photo that captures you without you knowing.
- Question 3: It is a good idea to protect strangers' privacy in your photos.
- Question 4: It is sometimes hard to avoid including strangers' faces in photos taken in public places.
- Question 5: If there was an option to protect a stranger's face in your photo without affecting quality, you would use it.
- Question 6: If there was an option to protect a stranger's face in your photo while slightly degrading the quality, you would still use it.
- Question 7: If there was an option to protect a stranger's face in your photo while significantly degrading the quality, you would still use it.

Users were asked to rate their opinions to these questions on a Likert scale.

Another portion of the survey presented an unaltered photo compared with anonymized photos where strangers' faces had been obfuscated using each of the three proposed methods. Participants were asked to rate their opinion on how harmful each method was to the original photo on a sliding scale from 0 to 10, with 0 in this case being not harmful and 10 being extremely harmful. In addition to rating the impact of each method on the photos, participants were also asked to rate their willingness to use each method on their own photos from 0 to 10. At the end of the section, participants could also optionally respond with testimonial as to which method they preferred and why.

In another portion of the survey, participants were presented with timed views of images. One image had strangers anonymized with face swapping and the other did not. The participants had to guess if any face swapping occurred or not in each of the two. This portion of the survey was intended to examine

how noticeable face swapping could be in a fast browsing environment such as social media where average users generally only spend a few seconds looking at pictures before moving on.

**Responses.** In total, the survey received 89 anonymous responses over the course of 1 month. Participants were primarily recruited among university students mainly majoring in computer science and computer engineering although a minority of respondents were working adults. All respondents participated on a completely voluntary basis (no incentives were provided). Exact demographic information was not collected to preserve participant anonymity. The detailed results for questions 1 through 7 are shown in Appendix B Fig. 9. In reviewing the responses to questions 1 through 7, it is clear that respondents had strong feelings in support of digital photo privacy. Most respondents likewise felt that both their privacy and stranger's privacy should ideally be protected in public photographs. Additionally, a large majority of respondents answered positively that they would make use of an anonymization system, assuming the impact to the photo was negligible while a majority responded negatively to any sort of significant impact to photo quality. Clearly, finding obfuscating methods with as little impact as possible to photo quality is paramount in designing a system that would be well-regarded and actually used.

The results of the harmfulness ratings questions are shown in Appendix B Fig. 10. The results of the willingness ratings questions are shown in Appendix B Fig. 11. From these results, black boxing received the most negative feedback with a large majority of participants rating its impact the worst overall and usability the lowest. Interestingly, blurring seemed to score the highest among respondents for both usability and harmfulness. Face swapping had comparable harmfulness, but was rated significantly lower on average for usability. To further examine these results, some participant responses provide helpful insight. Most participants who felt blurring was their preferred method seemed to find that face swapping was either unnatural looking, or felt that they could not trust it to always create a believable swap. For example, one user responded, "The black box hurt the quality of the photo and the swapping was disturbing because you could tell it was the wrong face on the stranger.". Overall, it seemed that blurring would be the best approach from a usability and perceived impact perspective based on this participant feedback.

Analyzing the number of users who were able to detect face swapping in the final survey section demonstrates that face swapping certainly is still detectable among many users despite recent advances in realistic swapping technology. 53.9% of respondents were able to tell that face swapping was used in the photo they were presented compared with 8% detection for the control photo. Many users noticed that something was different about the photo compared with the control photo, although the detection responses were close to random guessing for the photo with swapping.

## 8 Conclusion and Future Work

In this work, we presented a novel approach for automating the detection and anonymization of bystanders in digital photos, applying both a feature-based model and a privacy-concerned convolutional neural network. Techniques for feature engineering were explored with methods for utilizing metrics such as relative face size and head-pose estimation. Our MLP model achieved the highest validation accuracy at 94.34%, which demonstrates generalizability and promise for future use in an anonymization system for smartphone users. The convolutional neural network also demonstrated promising results with the highest achieved accuracy of 81.55% and limited overfitting of the training set. This work is the first of its kind in being able to offer a fully privacy-concerned approach as all other works previously relied upon contextual information within a full image. Additionally, we eliminate the need for any sort of manual cooperation between photographer and bystanders as most other related works require. The hope is that being able to offer a system that automates the protection of individuals in mobile photos and preserves the privacy of those captured, the user trust and willingness to use such a method in a real-world system is greatly enhanced over any previous methods which all require some form of participation on the bystanders' parts.

Three fully automated approaches for face anonymization (black boxing, blurring, and face swapping) were presented for use with the classifying models. To better understand user opinions around public photo privacy and each of the presented methods, a comprehensive user study was carried out. Participant responses indicated that while privacy of photos and individuals in public settings was definitely a concern for most, developing anonymizing methods which do not harm photo quality is important for creating any sort of real world system. Especially promising were the large number of positive responses on the blurring and swapping methods which indicate that the system has attraction to real-world users.

Due to the lack of available image datasets for the target/bystander detection application, collecting usable images including a good diversity of photo types, locations, lighting, etc. was a time-consuming process that could hardly be fully automated as photos had to be manually reviewed for suitability. The induced small size of the dataset might limit the generality of the results. In the future, we will expand upon this dataset to greatly increase the number of images and faces included. The goal of this is to ensure that models trained over the expanded dataset will be able to better generalize to in-the-wild photos. Making the dataset publicly available will allow other researchers to contribute and refine the images as well. Additionally, we plan to carry out a more significant user study with a larger population to examine the usability of our system in greater detail.

**Acknowledgements.** We thank Murtuza Jadliwala for shepherding this paper. We also thank the anonymous reviewers for their valuable comments.

# Appendix

## A Performance Characteristics of Classifiers

**Table 3.** Average single prediction forward-pass runtime (Intel i9-10900k)

| Learning algorithm | Average runtime (ms) |
|---|---|
| MLP | 0.0743 |
| SVM | 0.0541 |
| RF | 0.0515 |
| GDBT | 0.2028 |
| CNN | 244.3 |

Table 3 shows the measured single forward-pass runtimes for each of the examined classifiers averaged over 1000 runs. All feature-based classifiers have almost negligible runtime requirements for prediction operations on an Intel i9 platform. This indicates they would be excellent candidates to run directly on resource-constrained mobile devices.

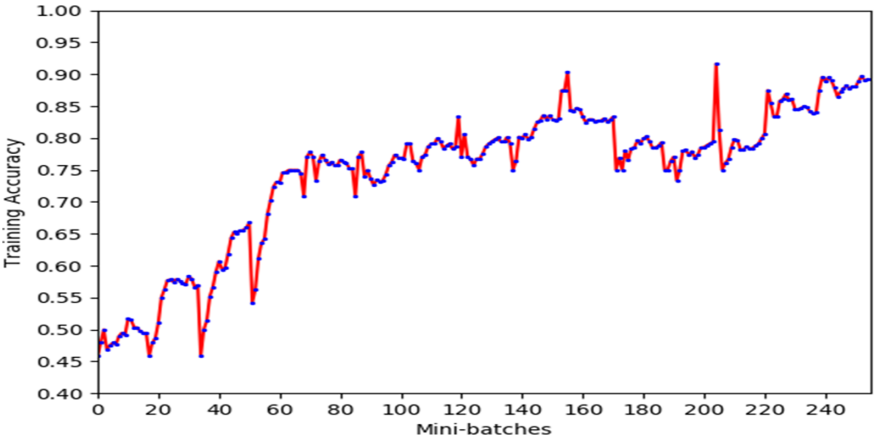


**Fig. 8.** Progressive accuracy over training mini-batches for the CNN classifier.

Figure 8 shows the accuracy of the CNN as mini-batches progress during training. Taken in conjunction with Fig. 6, the CNN appears to only converge after the 220th mini-batch where loss and accuracy variance is lowest.

## B Detailed Survey Results

Figure 9 shows detailed results of participant responses to survey questions 1–7. It is clear from these results that almost all respondents believe that photo privacy is important. Additionally, the rapid transition in willingness to use a system that does not degrade photo quality shows how paramount developing minimally invasive obfuscation methods is for real-world use.

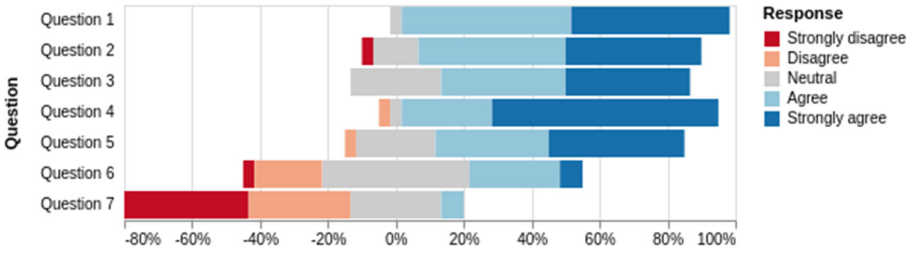


**Fig. 9.** Survey responses to opinion questions 1–7.

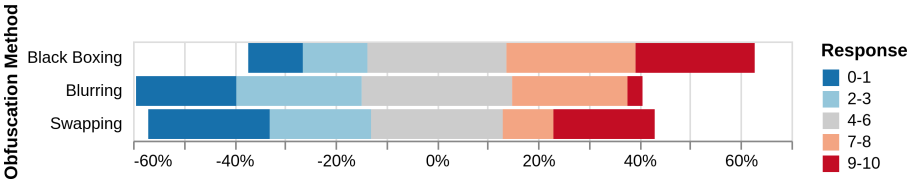


**Fig. 10.** Survey responses to rating the impact of anonymization methods on photos.

Figure 10 shows that respondents generally believe that the blurring and swapping methods are the least impactful to the sample photos. Interestingly, there was not a strong correlation between responses here and with the next set of questions over user willingness to use these same methods on their own photos.

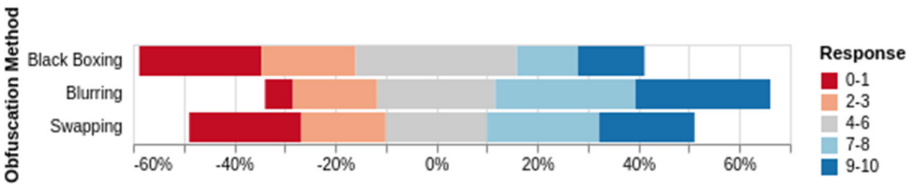


**Fig. 11.** Survey responses to rating how willing users would be to use each anonymization method.

The responses on willingness to use the obfuscation methods (see Fig. 11) demonstrate that there are more factors in determining what users want to use than perceived impact to photo quality. For example, face swapping, while rated as generally having a low impact to photo quality, received many negative responses from users in their likelihood to actually use it.

## References


1. Baltrusaitis, T., Robinson, P., Morency, L.: Constrained local neural fields for robust facial landmark detection in the wild. In: 2013 IEEE International Conference on Computer Vision Workshops, pp. 354–361, December 2013. https://doi.org/10.1109/ICCVW.2013.54
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.: OpenFace 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp. 59–66, May 2018. https://doi.org/10.1109/FG.2018.00019
3. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2930–2940, December 2013. https://doi.org/10.1109/TPAMI.2013.23
4. Bo, C., Shen, G., Liu, J., Li, X.Y., Zhang, Y., Zhao, F.: Privacy.tag: privacy concern expressed and respected. In: Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems. (SenSys 2014), pp. 163–176. Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2668332.2668339
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
6. Darling, D.: Target bystander detection repository (2020). https://github.com/ddarling/target-bystander-detection
7. Darling, D., Li, A., Li, Q.: Feature-based model for automated identification of subjects and bystanders in photos. In: IEEE International Workshop on the Security, Privacy, and Digital Forensics of Mobile Systems and Networks (MobiSec) (2019)
8. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D face reconstruction and dense alignment with position map regression network. In: ECCV (2018)
9. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002). https://doi.org/10.1016/S0167-9473(01)00065-2, http://www.sciencedirect.com/science/article/pii/S0167947301000652, nonlinear Methods and Data Mining
10. Hasan, R., Crandall, D., Fritz, M., Kapadia, A.: Automatically detecting bystanders in photos to reduce privacy risks. In: IEEE Symposium on Security and Privacy (S and P), May 2020. https://publications.cispa.saarland/3051/
11. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. **13**(4), 18–28 (1998). https://doi.org/10.1109/5254.708428
12. Ilia, P., Polakis, I., Athanasopoulos, E., Maggi, F., Ioannidis, S.: Face/off: preventing privacy leakage from photos in social networks. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. (CCS 2015), pp. 781–792. Association for Computing Machinery, New York (2015). https://doi.org/10.1145/2810103.2813603

13. Jung, J., Philipose, M.: Courteous glass. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. (UbiComp 2014), pp. 1307–1312. Adjunct, Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2638728.2641711
14. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
15. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision - ECCV 2012, pp. 679–692. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_49
16. Li, A., Darling, D., Li, Q.: PhotoSafer: content-based and context-aware private photo protection for smartphones. In: IEEE Symposium on Privacy-Aware Computing (PAC), pp. 10–18 (2018)
17. Li, A., Du, W., Li, Q.: PoliteCamera: respecting strangers' privacy in mobile photographing. In: 2018 International Conference on Security and Privacy in Communication Networks (SecureComm) (2018)
18. Li, A., Li, Q., Gao, W.: PrivacyCamera: cooperative privacy-aware photographing with mobile phones. In: IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 1–9 (2016)
19. Li, F., Sun, Z., Li, A., Niu, B., Li, H., Cao, G.: HideMe: privacy-preserving photo sharing on social networks. In: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, pp. 154–162, April 2019. https://doi.org/10.1109/INFOCOM.2019.8737466
20. Pech-Pacheco, J.L., Cristobal, G., Chamorro-Martinez, J., Fernandez-Valdivia, J.: Diatom autofocusing in brightfield microscopy: a comparative study. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 3, pp. 314–317, September 2000. https://doi.org/10.1109/ICPR.2000.903548
21. Raval, N., Srivastava, A., Lebeck, K., Cox, L., Machanavajjhala, A.: Markit: privacy markers for protecting visual secrets. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. (UbiComp 2014), pp. 1289–1295. Adjunct, Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2638728.2641707
22. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement (2018)
23. Richter, F.: Infographic: smartphones cause photography boom, August 2017. https://www.statista.com/chart/10913/number-of-photos-taken-worldwide/
24. Schiff, J., Meingast, M., Mulligan, D.K., Sastry, S., Goldberg, K.: Respectful cameras: detecting visual markers in real-time to address privacy concerns. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 971–978, October 2007. https://doi.org/10.1109/IROS.2007.4399122
25. Wood, E., Baltruaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3756–3764, December 2015. https://doi.org/10.1109/ICCV.2015.428
26. Xu, K., Guo, Y., Guo, L., Fang, Y., Li, X.: My privacy my decision: control of photo sharing on online social networks. IEEE Trans. Dependable Secure Comput. **14**(2), 199–210 (2017). https://doi.org/10.1109/TDSC.2015.2443795
27. Zadeh, A., Baltrušaitis, T., Morency, L.: Convolutional experts constrained local model for facial landmark detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2051–2059 (2017)