Automatically Detecting Bystanders in Photos to Reduce Privacy Risks

Rakibul Hasan[1], David Crandall[1], Mario Fritz[2], Apu Kapadia[1]
[1]Indiana University, Bloomington, USA
{rakhasan, djcran, kapadia}@indiana.edu
[2]CISPA Helmholtz Center for Information Security
Saarland Informatics Campus, Germany
fritz@cispa.saarland

***Abstract*—Photographs taken in public places often contain bystanders – people who are not the main subject of a photo. These photos, when shared online, can reach a large number of viewers and potentially undermine the bystanders' privacy. Furthermore, recent developments in computer vision and machine learning can be used by online platforms to identify and track individuals. To combat this problem, researchers have proposed technical solutions that require bystanders to be proactive and use specific devices or applications to broadcast their privacy policy and identifying information to locate them in an image.**

We explore the prospect of a different approach – identifying bystanders solely based on the visual information present in an image. Through an online user study, we catalog the rationale humans use to classify subjects and bystanders in an image, and systematically validate a set of intuitive concepts (such as intentionally *posing* for a photo) that can be used to automatically identify bystanders. Using image data, we infer those concepts and then use them to train several classifier models. We extensively evaluate the models and compare them with human raters. On our initial dataset, with a 10-fold cross validation, our best model achieves a mean detection accuracy of 93% for images when human raters have 100% agreement on the class label and 80% when the agreement is only 67%. We validate this model on a completely different dataset and achieve similar results, demonstrating that our model generalizes well.

***Index Terms*—privacy, computer vision, machine learning, photos, bystanders**

I. Introduction

The ubiquity of image capturing devices, such as traditional cameras, smartphones, and life-logging (wearable) cameras, has made it possible to produce vast amounts of image data each day. Meanwhile, online social networks make it easy to share digital photographs with a large population; e.g., more than 350 million images are uploaded each day to Facebook alone [1]. The quantity of uploaded photos is expected to only rise as photo-sharing platforms such as Instagram and Snapchat continue to grow [2], [3].

A large portion of the images shared online capture 'bystanders' – people who were photographed incidentally without actively participating in the photo shoot. Such incidental appearances in others' photos can violate the privacy of bystanders, especially since these images may reside in cloud servers indefinitely and be viewed and (re-)shared by a large number of people. This privacy problem is exacerbated by computer vision and machine learning technologies that can automatically recognize people, places, and objects, thus making it possible to search for specific people in vast image collections [4]–[6]. Indeed, scholars and privacy activists called it the 'end of privacy' when it came to light that Clearview – a facial recognition app trained with billions of images scraped from millions of websites that can find people with unprecedented accuracy and speed – was being used by law enforcement agencies to find suspects [7]–[9]. Such capabilities can easily be abused for surveillance, targeted advertising, and stalking that threaten peoples' privacy, autonomy, and even physical security.

Recent research has revealed peoples' concerns about their privacy and autonomy when they are captured in others' photos [10]–[12]. Conflicts may arise when people have different privacy expectations in the context of sharing photographs in social media [13], [14], and social sanctioning may be applied when individuals violate collective social norms regarding privacy expectations [15], [16]. On the other hand, people sharing photos may indeed be concerned about the privacy of bystanders. Pu and Grossklags determined how much, in terms of money, people value 'other-regarding' behaviors such as protecting others' information [17]. Indeed, some photographers and users of life-logging devices report that they delete photos that contain bystanders [18], [19], e.g., out of a sense of "propriety" [19].

A variety of measures have been explored to address privacy concerns in the context of cameras and bystanders. Google Glass's introduction sparked investigations around the world, including by the U.S. Congressional Bi-Partisan Privacy Caucus and Data Protection Commissioners from multiple countries, concerning its risks to privacy, especially regarding its impact on non-users (i.e., bystanders) [20], [21]. Some jurisdictions have banned cameras in certain spaces to help protect privacy, but this heavy-handed approach impinges on the benefits of taking and sharing photos [22]–[25]. Requiring that consent be obtained from all people captured in a photo is another solution but one that is infeasible in crowded places.

Technical solutions to capture and share images without infringing on other people's privacy have also been explored, typically by preventing pictures of bystanders from being taken or obfuscating parts of images containing them. For example, Google Street View [26] treats every person as a bystander

© 2020, Rakibul Hasan. Under license to IEEE.
DOI 10.1109/SP40000.2020.00097

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

and blurs their face, but this aggressive approach is not appropriate for consumer photographs since it would destroy the aesthetic and utility value of the photo [27], [28]. More sophisticated techniques selectively obscure people based on their privacy preferences [29]–[33], which are detected by nearby photo-taking devices (e.g., with a smartphone app that broadcasts preference using Bluetooth). Unfortunately, this approach requires the bystanders – the victims of privacy violations – to be proactive in keeping their visual data private. Some proposed solutions require making privacy preferences public (e.g., using visual markers [34] or hand gestures [33]) and visible to everyone, which in itself might be a privacy violation. Finally, these tools are aimed at preventing privacy violations as they happen and cannot handle the billions of images already stored in devices or the cloud.

We explore a complementary technical approach: automatically detecting bystanders in images using computer vision. Our approach has the potential to enforce a privacy-by-default policy in which bystanders' privacy can be protected (e.g., by obscuring them) without requiring bystanders to be proactive and without obfuscating the people who were meant to play an important role in the photo (i.e., the subjects). It can also be applied to images that have already been taken. Of course, detecting bystanders using visual features alone is challenging because the difference between a subject and a bystander is often subtle and subjective, depending on the interactions among people appearing in a photo as well as the context and the environment in which the photo was taken. Even defining the concepts of 'subject' and 'bystander' is challenging, and we could not find any precise definition in the context of photography; the Merriam-Webster dictionary defines 'bystander' in only a general sense as "one who is present but not taking part in a situation or event: a chance spectator," leaving much open to context as well as social and cultural norms.

We approach this challenging problem by first conducting a user study to understand how people distinguish between subjects and bystanders in images. We found that humans label a person as 'subject' or 'bystander' based on social norms, prior experience, and context, in addition to the visual information available in the image (e.g., a person is a 'subject' because they were interacting with other subjects). To move forward in solving the problem of automatically classifying subjects and bystanders, we propose a set of high-level visual characteristics of people in images (e.g., willingness to be photographed) that intuitively appear to be relevant for the classification task and can be inferred from features extracted from images (e.g., facial expression [35]). Analyzing the data from this study, we provide empirical evidence that these visual characteristics are indeed associated with the rationale people utilize in distinguishing between subjects and bystanders. Interestingly, exploratory factor analysis on this data revealed two underlying social constructs used in distinguishing bystanders from subjects, which we interpret as 'visual appearance' and 'prominence' of the person in a photo.

We then experimented with two different approaches for classifying bystanders and subjects. In the first approach, we trained classifiers with various features extracted from image data, such as body orientation [36] and facial expression [35]. In the second approach, we used the aforementioned features to first predict the high-level, intuitive visual characteristics and then trained a classifier on these estimated features. The average classification accuracy obtained from the first approach was 76%, whereas the second approach, based on high-level intuitive characteristics, yielded an accuracy of 85%. This improvement suggests that the high-level characteristics may contain information more pertinent to the classification of 'subject' and 'bystander', and with less noise compared to the lower-level features from which they were derived. These results justify our selection of these intuitive features, but more importantly, it yields an intuitively-explainable and entirely automatic classifier model where the parameters can be reasoned about in relation to the social constructs humans use to distinguish bystanders from subjects.

II. Related work

Prior work on alleviating privacy risks of bystanders can be broadly divided into two categories – techniques to handle images i) stored in the photo-capturing device and ii) after being uploaded to the cloud (Perez et al. provide a taxonomy of proposed solutions to protect bystanders' privacy [37]).

A. Privacy protection in the moment of photo capture

1) Preventing image capture: Various methods have been proposed to prevent capturing photographs to protect the privacy of nearby people. One such method is to temporarily disable photo-capturing devices using specific commands which are communicated by fixed devices (such as access points) using Bluetooth and/or infrared light-based protocols [38]. One limitation of this method is the photographers would have to have compliant devices. To overcome this limitation, Truong *et al.* proposed a 'capture resistant environment' [39] consisting of two components: a camera detector that locates camera lenses with charged coupled devices (CCD) and a camera neutralizer that directs a localized beam of light to obstruct its view of the scene. This solution is, however, effective only for cameras using CCD sensors. A common drawback shared by these location-based techniques [38], [39] is that it might be infeasible to install them in every location.

Aditya et al. proposed I-Pic [29], a privacy enhanced software platform where people can specify their privacy policies regarding photo-taking (i.e., allowed or not to take photo), and compliant cameras can apply these policies over encrypted image features. Although this approach needs the active participation of bystanders, Steil *et al.* proposed PrivacEye [40], a prototype system to automatically detect and prevent capturing images of people by automatically covering the camera with a shutter. Although there is no action needed from the bystanders to protect their privacy, PrivacEye [40] considers every person appearing in an image, limiting its applicability in more general settings of photography.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

The main drawback with these approaches is that they seek to completely prevent the capture of the image. In many cases, this may be a heavy-handed approach where removing or obscuring bystanders is more desirable.

2) Obscuring bystanders: Several works utilize image-obfuscation techniques to obscure bystanders images, instead of preventing image capture in the first place. Farinella *et al.* developed FacePET [41] to protect facial-privacy by distorting the region of an image containing a face. It makes use of glasses to emit light patterns designed to distort the Haar-like features used in some face detection algorithms. Such systems, however, will not be effective for other face detection algorithms such as deep learning-based approaches. COIN [30] lets users broadcast privacy policies and identifying information in much the same way as I-Pic [29] and obscure identified bystanders. In the context of wearable devices, Dimiccoli *et al.* developed deep-learning based algorithms to recognize activities of people in egocentric images degraded in quality to protect the privacy of the bystanders [42].

Another set of proposed solutions enable people to specify privacy preferences *in situ*. Li *et al.* present PrivacyCamera [43], a mobile application that handles photos containing at most two people (either one bystander, or one target and one bystander). Upon detecting a face, the app sends notifications to nearby bystanders who are registered users of the application using short-range wireless communication. The bystanders respond with their GPS coordinates, and the app then decides if a given bystander is in the photo based on the position and orientation of the camera. Once the bystander is identified (e.g., the smaller of the two faces), their face is blurred. Ra *et al.* proposed Do Not Capture (DNC) [31], which tries to protect bystanders' privacy in more general situations. Bystanders broadcast their facial features using a short-range radio interface. When a photo is taken, the application computes motion trajectories of the people in the photo, and this information is then combined with facial features to identify bystanders, whose faces are then blurred.

Several other papers allow users to specify default privacy policies that can be updated based on context using gestures or visual markers. Using Cardea [32], users can state default privacy preferences depending on location, time, and presence of other users. These static policies can be updated dynamically using hand gestures, giving users flexibility to tune their preferences depending on the context. In a later work, Shu *et al.* proposed an interactive visual privacy system that uses tags instead of facial features to obtain the privacy preferences of a given user [33]. This is an improvement over Cardea's system since facial features are no longer required to be uploaded. Instead, different graphical tags (such as a logo or a template, printed or stuck on clothes) are used to broadcast privacy preferences, where each of the privacy tags refer to a specific privacy policy, such as 'blur my face' or 'remove my body'.

In addition to the unique limitations of each of the aforementioned techniques, they also share several common drawbacks. For example, solutions that require transmitting bystanders' identifying features and/or privacy policies over wireless connections are prone to Denial of Service attacks if an adversary broadcasts this data at a high rate. Further, there might not enough time to exchange this information when the bystander (or the photographer) is moving and goes outside of the communication range. Location-based notification systems might have limited functionality in indoor spaces. Finally, requiring extra sensors, such as GPS for location and Bluetooth for communication, may prevent some devices (such as traditional cameras) from adopting them.

B. Protecting bystanders' privacy in images in the cloud

Another set of proposed solutions attempts to reduce privacy risks of the bystanders after their photos have been uploaded to the cloud. Henne *et al.* proposed SnapMe [44], which consists of two modules: a client where users register, and a cloud-based watchdog which is implemented in the cloud (e.g., online social network servers). Registered users can mark locations as private, and any photo taken in such a location (as inferred from image meta-data) triggers a warning to all registered users who marked it as private. Users can additionally let the system track their locations and send warning messages when a photo is captured nearby their current location. The users of this system have to make a privacy trade-off, since increasing visual privacy will result in a reduction in location privacy.

Bo *et al.* proposed a privacy-tag (a QR code) and an accompanying privacy-preserving image sharing protocol [34] which could be implemented in photo sharing platforms. The preferences from the tag contain a policy stating whether or not photos containing the wearer can be shared, and if so, with whom (i.e. in which domains/PSPs). If sharing is not permitted, then the face of the privacy tag wearer is replaced by a random pattern generated using a public key from the tag. Users can control dissemination by selectively distributing their private keys to other people and/or systems to decrypt the obfuscated regions. More recently, Li and colleagues proposed HideMe [45], a plugin for social networking websites that can be used to specify privacy policies. It blurs people who indicated in their policies that they do not want to appear in other peoples' photos. The policies can be specified based on scenario instead of for each image.

A major drawback of these cloud-based solutions is that the server can be overwhelmed by uploading a large number of fake facial images or features. Even worse, an adversary can use someone else's portrait or facial features and specify an undesirable privacy policy. Another limitation is that they do not provide privacy protection for the images that were uploaded in the past and still stored in the cloud.

C. Effectively obscuring privacy-sensitive elements in a photo

After detecting bystanders, most of the work described above obfuscate them using image filters (e.g., blurring [43]) or encrypting regions of an image [46], [47]. Prior research has discovered that not all of these filters can effectively obscure the intended content [27]. Masking and scrambling regions of interest, while effective in protecting privacy, may result in

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

a significant reduction of image utility such as 'information content' and 'visual aesthetics' [27]. In the context of sharing images online, privacy-protective mechanisms, in addition to being effective, are required to preserve enough utility to ensure their wide adoption. Thus, recent work on image privacy has attempted to maximize both the effectiveness and utility of obfuscation methods [28], [48]. Another line of research focuses solely on identifying and/or designing effective and "satisfying" (to the viewer) image filters to obfuscate privacy-sensitive attributes of people (e.g., identify, gender, and facial expression) [27], [49]–[51]. Our work is complementary to these efforts and can be used in combination with them to first automatically identify *what to obscure* and then use the appropriate obfuscation method.

III. Study Method

We begin with an attempt to define the notions of 'bystander' and 'subject' specific to the context of images. According to general dictionary definitions,[1,2,3] a bystander is a person who is *present and observing* an event *without taking part* in it. But we found these definitions to be insufficient to cover all the cases that can emerge in photo-taking situations. For example, sometimes a bystander may not even be *aware of* being photographed and, hence, not observe the photo-taking event. Other times, a person may be the subject of a photo without actively participating (e.g., by posing) in the event or even noticing being photographed, e.g., a performer on stage being photographed by the audience. Hence, our definitions of 'subject' and 'bystander' are centered around *how important a person in a photo is* and *the intention of the photographer*. Below, we provide the definitions we used in our study.

Subject: A subject of a photo is a person who is important for the meaning of the photo, e.g., the person was captured intentionally by the photographer.

Bystander: A bystander is a person who is not a subject of the photo and is thus not important for the meaning of the photo, e.g., the person was captured in a photo only because they were in the field of view and was not intentionally captured by the photographer.

The task of the bystander detector (as an 'observer' of a photo) is then to infer the importance of a person for the meaning of the photo and the intention of the photographer. But unlike human observers, who can make use of past experience, the detector is constrained to use only the visual data from the photo. Consequently, we turned to identifying a set of visual characteristics or high-level concepts that can be directly extracted or inferred from visual features and are associated with human rationales and decision criteria.

A central concept in the definition of bystander is whether a person is actively participating in an event. Hence, we look for the visual characteristics indicating *intentional posing* for a photo. Other related concepts to this are *being aware* of the photo shooting event and *willingness* to be a part of it. Moreover, we expect someone to look *comfortable* while being photographed if they are intentionally participating. Other visual characteristics signal the *importance of a person for the semantics of the photo* and whether they were *captured deliberately* by the photographer. We hypothesize that humans infer these characteristics from context and the environment, location and size of a person, and interactions among people in the photo. Finally, we are also interested to learn how the photo's environment (i.e., a public or a private space) affect peoples' perceptions of subjects and bystanders.

To empirically test the validity of this set of high-level concepts and to identify a set of image features that are associated with these concepts that would be useful as predictors for automatic classification, we conducted a user study. In the study, we asked participants to label people in images as 'bystanders' or 'subjects' and to provide justification for their labels. Participants also answered questions relating to the high-level concepts described above. In the following subsections, we describe the image set used in the study and the survey questionnaire.

A. Survey design

1) Image set: We used images from the *Google open image dataset* [52], which has nearly 9.2 million images of people and other objects taken in unconstrained environments. This image dataset has annotated bounding boxes for objects and object parts along with associated class labels for object categories (such as 'person', 'human head', and 'door handle'). Using these class labels, we identified a set of 91,118 images that contain one to five people. Images in the Google dataset were collected from Flickr without using any predefined list of class names or tags [52]. Accordingly, we expect this dataset to reflect natural class statistics about the number of people per photo. Hence, we attempted to keep the distribution of images containing a specific number of people the same as in the original dataset. To use in our study, we randomly sampled 1,307, 615, 318, 206, and 137 images containing one to five people, respectively, totaling to 2,583 images. A 'stimulus' in our study is comprised of an image region containing a single person. Hence, an image with one person contributed to one stimulus, an image with two people contributed to two stimuli, and so on, resulting in a total of 5,000 stimuli. If there are N stimuli in an image, we made N copies of it and each copy was pre-processed to draw a rectangular bounding box enclosing one of the N stimuli as shown in Fig. 1. This resulted in 5,000 images corresponding to the 5,000 stimuli. From now on, we use the terms 'image' and 'stimulus' interchangeably.

2) Measurements: In the survey, we asked participants to classify each person in each image as either a 'subject' or 'bystander,' as well as to provide reasons for their choice. In addition to these, we asked to rate each person according to the 'high-level concepts' described above. Details of the survey questions are provided below, where questions 2 to 8 are related to the high-level concepts.

[1]https://www.merriam-webster.com/dictionary/bystander
[2]https://dictionary.cambridge.org/us/dictionary/english/bystander
[3]https://www.urbandictionary.com/define.php?term=bystander

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

(a) Image with a single person. (b) Image with five people where the stimulus is enclosed by a bounding box. (c) An image where the annotated area contains a sculpture.

Fig. 1. Example stimuli used in our survey.

1) **Which of the following statements is true for the person inside the green rectangle in the photo?** with answer options i) There is a person with some of the major body parts visible (such as face, head, torso); ii) There is a person but with no major body part visible (e.g., only hands or feet are visible); iii) There is just a depiction/representation of a person but not a real person (e.g., a poster/photo/sculpture of a person); iv) There is something else inside the box; and v) I don't see any box. This question helps to detect images that were annotated with a 'person' label in the original Google image dataset [52] but, in fact, contain some form of depiction of a person, such as a portrait or a sculpture (see Fig. 1). The following questions were asked only if one of the first two options was selected.
2) **How would you define the place where the photo was taken?** with answer options i) A public place; ii) A semi-public place; iii) A semi-private place; iv) A private place; and v) Not sure.
3) **How strongly do you disagree or agree with the following statement: The person inside the green rectangle was aware that s/he was being photographed?** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*.
4) **How strongly do you disagree or agree with the following statement: The person inside the green rectangle was actively posing for the photo.** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*.
5) **In your opinion, how comfortable was the person with being photographed?** with a 7-point Likert item ranging from *highly uncomfortable* to *highly comfortable*.
6) **In your opinion, to what extent was the person in the green rectangle unwilling or willing to be in the photo?** with a 5-point Likert item ranging from *completely unwilling* to *completely willing*.
7) **How strongly do you agree or disagree with the statement: The photographer deliberately intended to capture the person in the green box in this photo?** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*.
8) **How strongly do you disagree or agree with the following statement: The person in the green box can be replaced by another random person (similar looking) without changing the purpose of this photo.** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*. Intuitively, this question asks to rate the 'importance' of a person for the semantic meaning of the image. If a person can be replaced without altering the meaning of the image, then s/he has less importance.
9) **Do you think the person in the green box is a 'subject' or a 'bystander' in this photo?** with answer options i) Definitely a bystander; ii) Most probably a bystander; iii) Not sure; iv) Most probably a subject; and v) Definitely a subject. This question was accompanied by our definitions of 'subject' and 'bystander'.
10) Depending on the response to the previous question, we asked one of the following three questions: i) **Why do you think the person in the green box is a subject in this photo?** ii) **Why do you think the person in the green box is a bystander in this photo?** iii) **Please describe why do you think it is hard to decide whether the person in the green box is a bystander or a subject in this photo?** Each of these questions could be answered by selecting one or more options that were provided. We curated these options from a previously conducted pilot study where participants answered this question with free-form text responses. The most frequent responses in each case were then provided as options for the main survey along with a text box to provide additional input in case the provided options were not sufficient.

3) Survey implementation: The 5,000 stimuli selected for use in the experiment were ordered and then divided into sets of 50 images, resulting in 100 image sets. This was done such that each set contained a proportionally equal number of stimuli of images containing one to five people. Each survey participant was randomly presented with one of the sets, and each set was presented to at least three participants. The survey was implemented in Qualtrics [53] and advertised on Amazon Mechanical Turk (MTurk) [54]. It was restricted to MTurk workers who spoke English, had been living in the USA for at least five years (to help control for cultural variability [55]), and were at least 18 years old. We further required that workers have a high reputation (above a 95% approval rating on at least

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

1,000 completed HITs) to ensure data quality [56]. Finally, we used two attention-check questions to filter out inattentive responses [57] (see Appendix F).

4) Survey flow: The user study flowed as follows:

1. Consent form with details of the experiment, expected time to finish, and compensation.
2. Instructions on how to respond to the survey questions with a sample image and appropriate responses to the questions.
3. Questions related to the images as described in Section III-A2 for fifty images.
4. Questions on social media usage and demographics.

B. Survey participants and dataset labels

1) Demographic characteristics of the participants: Before performing any analysis, we removed data from 45 participants who failed at least one of the attention-check questions. This left us with responses from 387 participants. Of these, 221 (57.4%) identified themselves as male and 164 as female. One hundred and eighty nine (48.8%) participants fell in the age range of 30–49 years, followed by 154 (39.8%) aged 18–29 years. A majority of the participants identified as White (n=242, 62.5%) followed by 82 (21%) as Asian, and 20 (5%) as African American. One hundred and ninety one (49.3%) had earned a Bachelor's degree, and 71 (18.3%) had some college education. Most of the participants had at least one social media account (n=345, 89.1%), among which only 7% (n=30) indicated that they never share images on those media. Each participant was paid $7, which was determined through a pilot study where participants were also asked whether they considered the compensation to be fair. Participants were able to pause this survey and resume at a later time, as indicated by the long completion time (> 10 hours) for many of the participants. Therefore we analyzed the response times for the top quartile, which completed the survey in an average of 41 minutes. Thus we estimated that our compensation was in the range of $10/hour for the work on our survey.[4]

2) Final set of images and class labels: For each image, we collected responses from at least three participants. Next, we excluded data for any image for which at least two participants indicated that there was no person in that image (by responding with any one of the last three options for the first question as described in Section III-A2). This resulted in the removal of 920 images, and the remaining 4,080 images were used in subsequent analyses.[5] The class label of a person was determined using the mean score for question 9: a positive score was labeled as 'subject', a negative score was labeled as 'bystander', and zero was labeled as 'neither'. In this way, we found 2,287 (56.05%) images with the label 'subject', 1,515 (37.13%) with 'bystander', and 278 (6.8%) with 'neither'. In this paper, we concentrate on the binary classification task ('subject' and 'bystander') and exclude the images with the 'neither' label. In this final set of images, we have 2,287 (60.15%) 'subjects' and 1,515 (39.85%) 'bystanders'.

3) Feature set: As described in section III-A2, we asked survey participants to rate each image for several 'high-level concepts' (questions 2–8). The responses were converted into numerical values – the 'neutral' options (such as 'neither disagree nor agree') were assigned a zero score, the left-most options (such as 'strongly disagree') were assigned the minimum score (-3 for a 7-point item), and the right-most options (such as 'strongly agree') were assigned the maximum score (3 for a 7-point item). Then, for each image, the final value of each concept was determined by computing the mean of the coded scores across the participants. In addition to these, we calculated three other features using the annotation data from the original Google image dataset [52]: size and distance of a person and the total number of people in an image. We estimated the size of a person by calculating the area of the bounding box enclosing the person normalized by total area of the image. The distance refers to the Euclidean distance between the center of the bounding box and the center of the image and can be treated as the 'location' of a person with respect to the image center. Finally, by counting the number of bounding boxes for each image, we calculated the total number of people in that image. We combined these three features with the set of high-level concepts and refer to this combined set simply as 'features' in the subsequent sections.

IV. Method of analysis

To understand how humans classify 'subjects' and 'bystanders' in an image, first, we catalog the most frequently used reasons for the classification (from responses to question 10). Next, we quantify if and how much these reasons are associated with the features as detailed in section III-B3. Significant association would indicate the relevance of the 'high-level concepts' in distinguishing bystander and subject by humans, and serve as a validation for incorporating those concepts in the study. Then, we conducted regression analyses to measure how effective each of the features *individually* are in classifying subject and bystander. Finally, we conducted exploratory factor analysis (EFA) on the whole feature set to surface any underlying constructs that humans use in their reasoning. EFA also helped to group correlated features under a common factor (based on the absolute values of factor loadings), facilitating the selection of a subset of uncorrelated features. Informed by the regression and factor analyses, we identified multiple subsets of features to use as predictors in training classifiers. In the following subsections, we explain each of these steps in more detail.

A. Quantifying association between human reasoning and features

We employed Spearman's ρ, which measures the monotonic association between two variables as a correlation measure between the binarized reasons and the real-valued features [58]. Then, for each reason, we grouped the feature values based on whether this reason was used for classification and measured

[4]A more conservative estimate yielded about $8/hour for the top 50%, which took an average of 53 minutes.

[5]One of the authors manually checked these images and found that only 9 (0.9%) of them contained people.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

the average of the feature-values in those two groups. We computed Cohen's d (i.e., the standardized mean difference or 'effect-size') between the two groups and conducted significance tests. A significant difference between the means would signal a feature is indicative of a particular reason.

B. Measuring predictive-power of individual feature and selecting subset of uncorrelated features

We trained one logistic regression model for each feature (as predictor) to classify 'subject' and 'bystander'. The predictive power of each feature, i.e., how well it alone can predict the class label was assessed by interpreting the model parameters. Our eventual goal is to find a subset of features with (collectively) high predictive power but minimal correlation among them since correlated features can render the model unstable [58]. To find a subset of features that are minimally correlated among themselves but retains maximum variance of the outcome variable, we conducted exploratory factor analysis (EFA) which attempts to discover underlying factors of a set of variables. Below we outline the steps we followed while conducting the factor analysis.

- **Removing collinear variables**. Multiple collinear variables can unduly inflate the variance of each other (i.e. inflate contribution of the variables toward a factor) and so collinear variables should be removed before conducting EFA [59]. First, we standardized the features to remove structural multi-collinearity [60]. Then we tested for multicollinearity using 'variance inflation factor' (VIF). We removed features with VIF greater than five [58].
- **Determining the number of factors to extract**. We conducted principal component analysis (PCA) to estimate the amount of variance retained by each component. We decided the number of factors to extract from EFA using a scree plot [58], [59], [61].
- **Extracting and rotating factors**. After removing collinear variables and deciding on the number of factors, we extracted the factors and estimated the factor loading (i.e., correlation between a feature and a factor) of each feature. Finally, we rotated the factors using 'varimax' rotation to obtain a simple structure of the factor loadings [59], [61]. The factors become orthogonal (i.e. completely uncorrelated) to each other after the rotation, which makes interpretation easier. Moreover, it helps to group and describe the features, since ideally each feature has a high factor loading for only one factor after the rotation.

Features that are highly correlated among themselves measure the same underlying concept (i.e., factor) and would have high correlation with that factor. Consequently, we grouped the features having high correlation with a single factor into categories describing 'meaningful' constructs. This would facilitate in explaining the underlying constructs that are important in the human reasoning process [59]. Additionally, features belonging to one group ideally have low correlation with features belonging to another group. Thus, we identified a subset of minimally correlated features by taking one feature from each group. The collective predictive power of this subset is indicated by how much of the total variance in the full set of variables is retained by the factors.

C. Developing classifiers using selected feature sets

So far, we have detailed the methods of validating our feature set and identifying subsets of features to be used as predictors. Now, we focus on developing machine learning (ML) models and evaluating their performance. Although we strive to achieve high classification accuracy, we are also interested in learning at what level of abstraction the features have the most predictive power. Thus, we built several classifiers using features at different levels of abstraction, spanning from the raw image to the high-level concepts and evaluated these models by conducting 10-fold cross-validations. Below, we explain these different classifier models.

1) Baseline models: As a baseline model, we started with directly using the cropped images as features to train the classifier. All the cropped images were first resized (256×256 pixels) and then fed into a logistic regression model. This represents a model trained with the most concrete set of features, i.e., the raw pixel values of the cropped images. Our next classifier is another logistic regression model, trained with higher-level but simple features – the number of people in a photo and the size and the location of each person. This would allow us to investigate if the classification problem can be trivially solved using easily obtainable, simple features.

2) Fine-tuning pre-trained models: Fine-tuning a pre-trained model allows us to transfer learned knowledge in one task to perform some other (often related) task. The process is analogous to how humans use knowledge learned in one context to solve a new problem. Fine-tuning deep learning models has shown great promise in many related problem domains [62]–[65]. Here, we fine-tuned ResNet50 [66], which was trained for object detection and recognition on the ImageNet [67] dataset containing more than 14 million images to classify 'subject' and 'bystander'. We chose to use this model since recognizing an object as a 'person' is a pre-requisite to classify them as 'subject' or 'bystander'. Hence, the model parameters were pre-trained to optimize recognizing people (and other objects), and we fine-tune it to classify detected people as 'subject' or 'bystander'. To fine-tune this model, we replaced the final layer with a fully connected layer with 'sigmoid' activation function. This modified network was re-trained using our (cropped) image dataset. In fine-tuning, we only update the parameters of the last (i.e., newly added) layer, keeping the parameters of all the other layers intact.

3) Models with higher level features: In section IV-B, we outlined the process of examining the predictive power of the features and discovering a set of minimally correlated features that best predicts the outcome variable. The feature set includes the high-level concepts, which are not, unfortunately, directly derivable from the image data with currently available machine learning models. We attempt to overcome this barrier by utilizing existing ML models to extract features that we believe to be good proxies for the high level concepts. We then

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

train two classifiers by – 1) training directly with these proxy features and 2) following a *two-step* classification pipeline by first training regression models with the proxy features to predict the high-level concepts and then using the *predicted* values of the high-level concepts to train the final classifier. Below, we detail what proxy features we extracted and how.

- **Human related features.** The ResNet50 [66] model was trained to categorize objects (including people) in images. We feed the cropped images of people in our dataset in the pre-trained model and extract the output of the second-to-last layer of the network to be used as features for our classifier. Since the original RestNet50 network uses these features in the last layer to assign an object to the appropriate class, and the class in our case is 'person', the features are presumably useful in distinguishing people from other objects. In other words, these features are useful in detecting people, which is a prerequisite for classifying a person as a subject or bystander.
- **Body-pose related features.** We used OpenPose [36] to estimate body-pose of a person, which attempts to detect 18 regions (or joints) of a human body (such as nose, ears, and knees), and outputs detected joints along with detection confidence. We used the confidence scores, which indicate how clearly different body parts of a person are visible in an image, as feature values. Additionally, for each pair of neighboring joints (e.g., right shoulder and right elbow), we computed the angle between a line connecting these joints and the horizontal axis. Collectively, these angles suggest the pose and the orientation of the body. These features were extracted from OpenPose [36] using the cropped images of each person. But in our dataset, some cropped images contain body parts of more than one person (see Fig. 2), and OpenPose attempts to detect all of them. Since in our case a single stimulus (i.e. cropped image) is associated with one person, we needed to single out the pose features for that person only. For example, Fig. 2a shows a cropped image where two people are visible, but the original image was cropped according to the bounding box for the person at the right side of the cropped image. Although OpenPose detects body parts for both people, we need this information only for the person with whom this image is associated (in this case the person at the right side), since the pose features will be used to classify that person only. We use a simple heuristic to solve this problem – a cropped image is associated with the most centrally-located person. With this heuristic, when a body part (such as nose) was detected more than once, we retain information about the part that is closest to the center of the cropped image. Fig. 2b shows the result of body part detection using this mechanism.
- **Emotion features estimated from facial expression.** We extracted scores for seven emotions: 'angry', 'disgusted', 'fearful', 'happy', 'sad', 'surprised', and 'neutral'. Intuitively, these features might be good proxies for 'awareness', 'comfort', and 'willingness' of a person. To obtain emotion features, we first extracted faces from the cropped images using a face detection model [68]. If two people appear in each other's cropped images, each of them will be positioned in a more central location of the cropped image associated with them and will be detected with higher accuracy and confidence by the face detection algorithm. Hence, in cases where a cropped image contains multiple people, we retained the face that was detected with the highest confidence. After detection, the faces were extracted and fed into a facial expression recognition model [35]. Using facial features, this model estimates the probabilities of each of the seven emotions. We used these probability values as features.

(a) The colored dots show the body joints of the two people originally detected.

(b) Result of removing duplicate body joints based on the distance from image center.

Fig. 2. Detecting and refining body joints.

D. Comparing ML models with humans

One way to investigate how well the ML models perform compared to humans is to compare how much human annotators agree among themselves with the model accuracy. Computing agreement statistics, however, require all annotators to label the same set of images, which is infeasible in this case. Hence, instead of agreement among the annotators, we computed what percentage of annotators agreed with the final class label of an image. Recall that the final class label was decided by taking the mean of the scores for 'subject' and 'bystander' (provided by the survey participants). For example, if two participants labeled someone as 'most probably a subject' (coded value = 1), and a third participant labeled that person as 'most probably a bystander' (coded value = -1), then the mean score is 0.3. Hence, the final label of that person would be 'subject', where 67% annotators agreed with this label. We grouped the images based on what percentage of the annotators agreed with its label. We then used these groups individually to train classifiers and test their performance for image sets with varying degrees of agreement.

E. Test dataset

We assessed the performance and robustness of the models created with the above-mentioned steps with 10-fold cross-validation using non-overlapping train-test splits of the Google

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

dataset [52]. To evaluate how well our approach generalizes to different datasets, we conducted additional analysis (using the model trained on the Google dataset) on an independent dataset consisting of 600 images sampled from the *Common Objects in COntext* (COCO) dataset [69]. COCO contains a total of 2.5 million labeled instances in 328,000 images of complex everyday scenes containing common objects in their natural context and has been used in numerous studies as a benchmark for object recognition and scene understanding. We randomly sampled roughly equal number of photos with one to five people totalling to 600 samples of individual person. Using this sample, survey data was collected and analyzed in the same way as explained above, but participants from the previous study were not allowed to take this survey. After pre-processing the survey data, we found that 354 (59%) and 246 (41%) people in the images were labeled as 'subject' and 'bystander', respectively.

V. Findings

A. How humans classify 'subjects' and 'bystanders'?

The most frequently used reasons for labeling a person as a 'subject' or a 'bystander' by the survey participants are shown in Tables I and II. For 'subjects', the top four reasons involve visual characteristics of the individual person under consideration (Table I). Intuitively, these reasons are related with the visual features we extracted from the images and collected using survey responses (we quantify these associations and present the results in the next section). For example, 'being in focus' with size and location of a person, 'taking a large space' with size, and 'being the only person' and 'activity of the person being the subject matter of the image' with importance of the person for the semantic of the image or if the person can be replaced without altering the semantic content. The last three reasons consider overall image context and visual similarities of the person in question with other people in the same image (Table I).

Similarly, the most frequently selected reason for labeling a person as a 'bystander' (Table II) is 'not focusing on the person', which is associated with the size and location of that person in the image. The second most frequent reason is 'caught by chance', which again relates to if that person is important for the image or can be replaced. Reasons 4 and 5 were chosen when participants thought no person was a subject of the image or there was no specific subject at all. The other reasons consider overall image content and visual similarity and interactions of the person in question with other people in the image (Table II). These results indicate that the human decision process for this classification task considers visual characteristics of the person in question (e.g. size) as well as other people in the image (e.g. interaction among people in the image). This process also involves understanding the overall semantic meaning of the image (e.g., someone was captured by chance and not relevant for the image) and background knowledge (e.g., if two people have similar visual features or are performing the same activity, then they should belong to the same class). Such rich inferential knowledge is not available in images. Since our ultimate goal is to build classifiers that only use the images as input, we investigate the relationships of the human rationale with visual features that can be extracted from the image.

TABLE I
Most frequent reasons found in the pilot study for classifying a person as a *Subject* and how many times each of them was selected in the main study.

#	Reason	Frequency
1	This photo is focused on this person.	5091
2	This photo is about what this person was doing.	4700
3	This is the only person in the photo.	2740
4	This person is taking a large space in the photo.	2425
5	This person was doing the same activity as other subject(s) in this photo.	2357
6	This person was interacting with other subject(s) in this photo.	1715
7	The appearance of this person is similar to other subject(s) of this photo.	1644

TABLE II
Most frequent reasons found in the pilot study for classifying a person as a *Bystander* and how many times each of them was selected in the main study.

#	Reason	Frequency
1	This photo is not focused on this person.	3553
2	This person just happened to be there when the photo was taken.	2480
3	The activity of this person is similar to other bystander(s) in this photo.	1758
4	Object(s) other than people are the subject(s) of this photo.	1644
5	Appearance of this person is similar to other bystanders in this photo.	1278
6	There is no specific subject in this photo.	849
7	This person is interacting with other bystander(s).	755
8	This person is blocked by other people/object.	567
9	Appearance of this person is different that other subjects in this photo.	537
10	The activity of this person is different than other subjects(s) in this photo.	466

B. Association between human-reasoning and the features

1) How well are the 'high-level concepts' and the 'features' associated with the reasons humans used?: The correlations between the features and the reasons for specific labels and the standardized differences between the means in feature values when a specific rationale was used or not used for labeling are presented in Tables III and IV.[6] Significant correlation coefficients and differences in group means suggest an association between the features and the rationales. As an example, the positive correlation coefficient of 0.19 indicates that when participants thought that *the photo was focused on a person*,

[6]Since the features are related to individual people and do not capture the interactions among people or the overall contexts of the images, we present results only for the reasons referring to individual persons.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

TABLE III
CORRELATION COEFFICIENTS AND EFFECT SIZES BETWEEN THE VISUAL FEATURES AND THE REASONS FOR CLASSIFYING A PERSON AS A *subject*. ALL COEFFICIENTS AND EFFECT-SIZES ARE SIGNIFICANT AT $p < .001$ LEVEL.

Feature	Spearman ρ	Cohen d
This photo is focused on this person.		
Awareness	0.17	0.36
Pose	0.19	0.42
Comfort	0.15	0.30
Willingness	0.15	0.30
Replaceable	-0.20	-0.39
Size	0.35	0.69
Distance	-0.29	-0.63
Number of people	-0.37	-0.82
This person is taking a large space in the photo.		
Awareness	0.11	0.22
Comfort	0.11	0.24
Willingness	0.12	0.25
Replaceable	-0.20	-0.43
Size	0.38	0.83
Distance	-0.19	-0.43
Number of people	-0.20	-0.44
This is the only person in this photo.		
Awareness	0.11	0.21
Pose	0.10	0.21
Replaceable	-0.12	-0.24
Size	0.27	0.65
Distance	-0.23	-0.47
Number of people	-0.61	-1.33

TABLE IV
CORRELATION COEFFICIENTS AND EFFECT SIZES BETWEEN THE VISUAL FEATURES AND THE REASONS FOR CLASSIFYING A PERSON AS A *bystander*. ALL COEFFICIENTS AND EFFECT-SIZES ARE SIGNIFICANT AT $p < .001$ LEVEL.

Feature	Spearman ρ	Cohen d
This photo is not focused on this person.		
Awareness	-0.25	-0.59
Pose	-0.31	-0.77
Comfort	-0.25	-0.49
Willingness	-0.26	-0.52
Replaceable	0.16	0.31
Photo place	-0.22	-0.52
Size	-0.20	-0.44
Distance	0.21	0.46
This person just happened to be there when the photo was taken.		
Awareness	-0.34	-0.70
Pose	-0.36	-0.72
Comfort	-0.19	-0.33
Willingness	-0.22	-0.41
Replaceable	0.27	0.50
Photo place	-0.24	-0.49
Size	-0.23	-0.37
Distance	0.13	0.26
This person is blocked by other people or object.		
Awareness	-0.15	-0.46
Pose	-0.17	-0.54
Comfort	-0.11	-0.29
Willingness	-0.12	-0.37
Replaceable	0.14	0.38

they also tended to agree more on the assertion that that person was *posing* for the photo. Similarly, the (standardized) difference between the means of the 'Posing' feature when participants used the reason *the photo was focused on that person* to label a person as a subject versus when they did not used that reason is 0.42.[7] This implies that being 'in-focus' of a photo is related to the concept of 'posing' for that photo. Associations among the other reasons and high-level concepts can be similarly interpreted.

2) Identifying subsets of uncorrelated features that are effective in distinguishing 'subject' and 'bystander': First, we trained separate classifier models for each feature as a predictor to asses how well each of them can individually distinguish between a 'subject' and a 'bystander'. We report the detailed results in Appendix A. In summary, all of the features (described in Section III-B3) were found to be significantly associated with the outcome (i.e., subject and bystander), but the magnitude of the predictive power varied across features. We also found that almost all pairs of features have medium to high correlations between them (Appendix B). Hence, we conducted EFA to discover uncorrelated feature sets.

As outlined in Section IV, first we calculated VIF to detect multicollinearity (Table IX). Among the features, 'Awareness' has the highest VIF of 5.8 (and a corresponding $R^2 > .8$ in the regression model), indicating that this feature can be predicted almost perfectly using a linear combination of other features. This is also apparent in the pairwise correlations among the features (see Appendix B), where 'Awareness' is highly correlated with most of the other features, making it redundant. Removal of this feature resulted in a drop of VIF for every other feature below 5, suggesting a reduction in multicollinearity in the system (re-calculated VIF are shown in the second column of Table IX).

With the remaining features, we conducted PCA to find out the appropriate number of factors to extract [59]. The point of inflexion [59] in the Scree plot (Fig. 3) after the second factor suggests the extraction of two factors, which jointly retain approximately 60% of the total variance in the data. Fig. 4 exhibits the factor loadings of each feature after a 'varimax' rotation [58]. We omitted the features with factor loadings less than 0.32 [61].[8] A feature is associated with the factor with which it has a higher loading than the other, and the features associated with the same factor were grouped together to form descriptive categories [59]. More specifically, 'Pose', 'Comfort', and 'Willingness' were grouped together under the

[7]Cohen's d=0.2, 0.5, and 0.8 are considered to be a 'small', 'medium', and 'large' effect size respectively [70].

[8]The location of a person did not have high enough correlation with any of the factors. Hence, it was not used in subsequent analysis.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

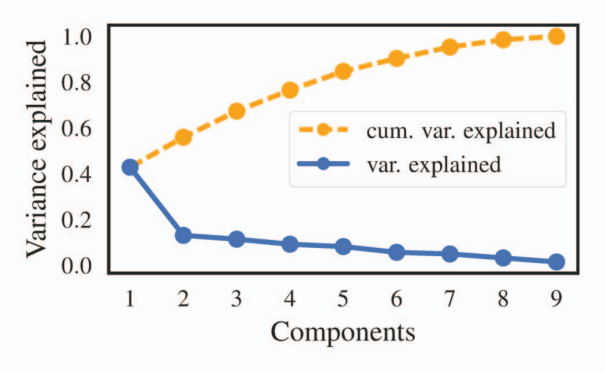


Fig. 3. Scree plot showing *proportions of variance* and *cumulative proportion of variance* explained by each component extracted using PCA.

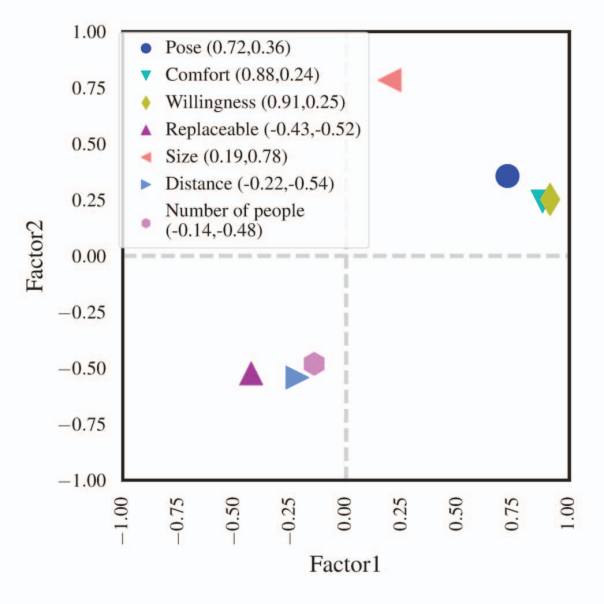


Fig. 4. Factor loadings of the features across the two extracted factors. The numeric values of the loadings are displayed within braces with the legend.

category 'visual appearance' of a person. This grouping makes sense intuitively as well since all three variables refer to the body orientation and facial expression of a person. Similarly, 'Size', 'Distance', and 'Number of people' collectively represent 'how prominent' the person is in the photo.[9] Finally, 'Replaceable' has almost equal loadings on the two factors and, hence, was not assigned to any group. Intuitively, it suggests how 'important' a person is for the semantic meaning of the image, which depends on both the 'visual appearance' and 'prominence' of a person.

Upon grouping the features that are highly correlated among themselves, we now select a subset of features by picking one feature from each group ('Pose' and 'Size', respectively) and the two features ('Replaceable', and 'Photographer's intention') that do not belong to any group.[10] 'Replaceable', and 'Photographer's intention'. Results from a linear regression model trained with this feature set is shown in Table V. This model has a better fit with the data ($R^2 = 0.53$) than any of the models trained with individual features (Table VII). But this model utilizes ground truth data about 'Pose', 'Replaceable', and 'Photographer's intention' obtained from the user study, which can not be extracted directly from the image data. In the next section, we present classification results using different feature sets produced from the images.

C. Machine learning models to predict 'subject' and 'bystander'

Table VI shows means and standard-deviations for classification accuracy using different feature sets (including the model using ground truth high-level concepts). Fig. 5 shows the corresponding Receiver Operating Characteristic (ROC) plots for each case generated from 10-fold cross-validation. Using the cropped images as features has the lowest mean accuracy of 66%. Using the simple features – 'Size', 'Distance', and 'Number of people' – yielded mean accuracy of

[9]Although 'Size' appears to be far from the others, this is because it has positive association with 'Factor2', while the rest have negative association. This is also intuitive, since as the 'Number of people' and 'Distance' increase, size should decrease.

[10]We experimented with different combinations of features from these two groups and obtained comparable results.

TABLE V
EFFECTIVENESS OF THE SELECTED FEATURES TO CLASSIFY 'SUBJECT' AND 'BYSTANDER'. THE COLUMNS SHOW ODDS-RATIOS AND THEIR 95% CONFIDENCE INTERVALS FOR EACH FEATURE. ALL $p < 0.0001$.

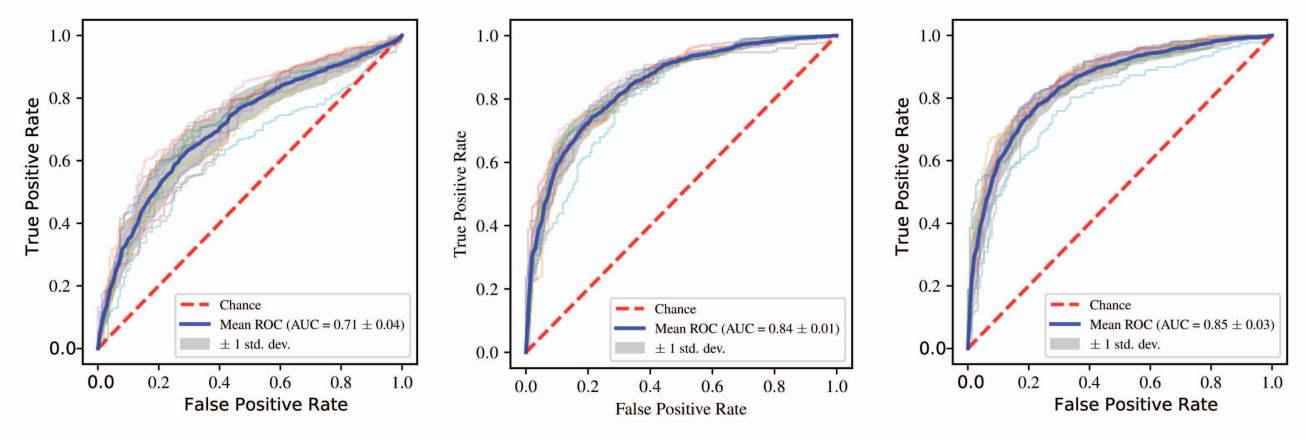
	Odds Ratio	[95% CI]
Pose	2.50	[2.17, 2.88]
Replaceable	0.13	[0.11, 0.15]
Size	1.91	[1.64, 2.22]
Photographer's intention	0.56	[0.49, 0.63]

76%, a 15% increase than using raw image data. We see a corresponding increase in the area under the curve (AUC) measure in Fig. 5. Fine-tuning the pre-trained ResNet [66] model did not improve the accuracy any further (Table VI).

Using ground truth values of the high-level concepts, combined with the 'Size' feature increased the accuracy by more than 12% (mean accuracy $86\% \pm 0.04$ and AUC 93%). Next, we employ the proxy features of these high-level concepts as detailed in Section IV-C3 and obtained a mean classification accuracy of 78%, a small increase from the model using simple features. Finally, we use the *predicted* values of the high-level concepts using the *proxy features* and obtained a mean accuracy of 85% and corresponding AUC of 93%, *which is similar to the results obtained using ground truth values* of the high-level concepts (details on prediction accuracy are provided in Appendix C). We obtained similar results using different subsets of predicted features, indicating that predictors in the same set contain repeated information and do not add any new predictive power, which again validates our EFA analysis.

From these results, we see that features at a higher level

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.



(a) Cropped image

(b) *Size*, *Distance*, and *Number of people*

(c) Fine-tuned *ResNet*

(d) *ResNet*, *OpenPose*, *Emotion*

(e) Ground truth *Pose*, *Replaceable*, *Photographer's intention*, and *Size*

(f) Predicted *Pose*, *Replaceable*, *Photographer's intention*, and *Size*

Fig. 5. Receiver operating characteristic (ROC) plots for classifier models using different feature sets.

TABLE VI
MEAN AND STANDARD DEVIATION OF ACCURACY FOR CLASSIFICATION USING DIFFERENT FEATURE SETS ACROSS 10-FOLD CROSS VALIDATION.

Features	Accuracy Mean	Accuracy SD
Cropped image	66%	0.03
Size, *distance*, and *number of people*	76%	0.01
Fine-tuning *ResNet*	77%	0.02
ResNet, *Pose*, and *Facial expression* features	78%	0.03
Size and ground truth *Pose*, *Replaceable*, *Photographer's intention*	86%	0.04
Size and predicted *Pose*, *Replaceable*, *Photographer's intention*	85%	0.02

of abstraction yield better classification accuracy. The raw image, despite having all the information present in any feature derived from it, performs noticeably worse than even the simple feature set. Similarly, *predicted* values of the high-level concepts performed better than the proxy features they were predicted from. Although the proxy features presumably contain more information than any feature predicted from them, the high-level concepts are more likely to contain information relevant for distinguishing subjects and bystanders in a more concise manner and with less noise.

D. Comparing ML models with humans

The percentages of agreement among the annotators and the number of images for each percentage are presented in Appendix D. All annotators agreed on the final label for only 1,309 (34%) images, and for 1,308 (34%) images there were agreements among two-third of the annotators. For these two groups of images, we train and evaluate classifiers following the two-step procedure.[11] For a 10-fold cross validation, the mean classification accuracy were $80\%(\pm 0.03)$ and $93\%(\pm 0.02)$, respectively for these two groups (The corresponding ROC plots are shown in Appendix E). Considering the fact that these two models were trained using much smaller sets of images than before, they achieved remarkably high accuracy even for the images with only 67% agreement among human annotators.

[11]We did not perform similar analyses for images with lower than 67% agreement because of insufficient training data. We had only 400 such images.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

E. Accuracy on the COCO dataset

For the 600 images sampled from COCO [69], our model (trained on the Google data set) achieved an overall classification accuracy of 84.3%. To compare the accuracy with humans, we again divided these images based on how many of the annotators agreed with the final label. We found that 354 (59%) images had 100% agreement, while 168 (28%) images had 67% agreement. For these two subsets, our model achieved 91.2% and 78.6% classification accuracy, respectively. The results of this extended analysis are consistent with the results with the Google dataset and provide strong evidence for the generalization of our approach and trained models.

VI. Limitations and Discussion

Photography as art. We must note that just because bystanders *can* be detected does not mean that they *should* be removed or redacted from images, or that a particular bystander should necessarily exert control over the image. There are legitimate reasons for bystanders to be retained in images, ranging from photo-journalism to art. The questions of image ownership and the right to privacy of bystanders are complicated and depend on contextual, cultural, and legal factors. Nevertheless, in many circumstances, owners of photos may voluntarily be willing to redact images out of a sense of 'propriety' and concern about bystanders [19]. For example, Anthony et al. discuss how people routinely engage in behaviors to respect the privacy of others [71]. Other work seeks to make privacy 'fun' by encouraging owners of photos to apply stickers or redactions on bystanders [27], [50]. Our work on detecting bystanders should thus be seen as a necessary building block of larger automated frameworks that consider further action on photos.

People detection. For the Google dataset [52], we used manually annotated bounding boxes to locate people and extracted features from these cropped images. Results may differ if people were instead detected automatically, but we do not expect large deviations since computer vision can detect and segment people with close to human-level performance [72].

Annotators. All of our survey participants were U.S. residents (although the images used had no such restriction); future work could consider cross-cultural studies. We used three annotators per image under the assumption that unanimous agreement among three independent observers is a strong signal that a given person is indeed a 'bystander' or 'subject'. We expect that requiring agreement among more annotators would slightly reduce the size of the dataset but also increase the accuracy of our algorithm for that dataset, as any ambiguity is further reduced. Overall, three annotators struck a reasonable balance for such labeling.

Dataset. We considered images containing one to five people for practical reasons. In our labeled data, we noticed that as the number of people per image grows, fewer of them are labeled as subjects. This indicates that, as one might expect, images with large numbers of people typically contain crowds in public places, with no particular subject. Including such images would result in an imbalanced dataset and ultimately a biased model.We hypothesize that classifying subjects and bystanders in such images would be easier than in images with fewer people since people usually have smaller size and are not centrally located (size and location features have significant positive and negative correlations with being a subject) in those images. Finally, we observed that beyond some threshold, people with smaller size are much harder to recognize. Thus, we expect that our algorithm will not only scale to images with larger crowds but will yield better classification accuracy.

Feature relationships. Another limitation of our work is that we use features only from individual people as predictors. However, as our user study uncovered, relationships and interactions among people in an image also play important roles in the categorization of *subject* vs. *bystander*. For example, some participants labeled a person as a 'bystander' because they "looked similar to" or "were doing the same activity as" another bystander. Future work should investigate classifiers that incorporate these inter-personal relationships.

Use of additional metadata. Our goal in this paper is to propose a general-purpose bystander detector using visual features alone, to make it as widely applicable as possible, including on social media platforms, image-hosting cloud servers, and photo-taking devices. We expect that accuracy can be increased using contextual information available in any specific domain, e.g., using image captions, one's friend list in a social network, and location of the photo. In the future, we plan to explore the use of domain-specific information.

VII. Conclusion

Photographs often inadvertently contain bystanders whose privacy can be put at risk by harming their social and professional personas. Existing technical solutions to detect and remove bystanders rely on people broadcasting their privacy preferences as well as identifying information – an undue burden on the victims of privacy violations. We attempt to tackle the challenging problem of detecting bystanders automatically so that they can be removed or obfuscated without proactive action. Our user study to understand the nuanced concepts of what makes a 'subject' vs. 'bystander' in a photo unveiled intuitive *high-level concepts* that humans use to distinguish between the two. With extensive experimentation, we discovered visual features that can be used to infer those concepts and assessed their predictive power. Finally, we trained machine learning models using selected subsets of those concepts as features and evaluated their performance. Our best classifier yields high accuracy even for the images in which the roles of subjects and bystanders are not very clear to human annotators. Since our system is fully automated, and solely based on image data, it does not require any additional setup and can be used for any past, present, and future images, we believe that it has the potential to protect bystanders' privacy at scale.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation under grant CNS-1408730. We also thank Ninja Marnau for her helpful comments.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] K. Smith, "53 Incredible Facebook Statistics and Facts," 2019. [Online]. Available: https://www.brandwatch.com/blog/facebook-statistics/

[2] N. Lomas, "Teens favoring Snapchat and Instagram over Facebook, says eMarketer," https://techcrunch.com/2017/08/22/teens-favoring-snapchat-and-instagram-over-facebook-says-emarketer/, 2017.

[3] I. A. Hamilton. (2019) Instagram has avoided Facebook's trust problem, beating its parent as app of choice for Generation Z. [Online]. Available: https://www.businessinsider.com/instagram-is-more-popular-among-generation-z-than-facebook-2019-3

[4] R. Shaw, "Recognition markets and visual privacy," *UnBlinking: New Perspectives on Visual Privacy in the 21st Century*, 2006.

[5] A. Acquisti, R. Gross, and F. D. Stutzman, "Face recognition and privacy in the age of augmented reality," *Journal of Privacy and Confidentiality*, vol. 6, no. 2, p. 1, 2014.

[6] M. Starr, "Facial recognition app matches strangers to online profiles," 2014. [Online]. Available: https://tinyurl.com/s58ytv8/

[7] K. Hill, "The Secretive Company That Might End Privacy as We Know It." 2020. [Online]. Available: https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

[8] M. Andrejevic and N. Selwyn, "Facial recognition technology and the end of privacy for good." 2020. [Online]. Available: https://lens.monash.edu/2020/01/23/1379547/facial-recognition-tech-and-the-end-of-privacy

[9] B. C. McCarthy and A. Feis, "Rogue NYPD cops are using facial recognition app Clearview." 2020. [Online]. Available: https://nypost.com/2020/01/23/rogue-nypd-cops-are-using-sketchy-facial-recognition-app-clearview/

[10] V. G. Motti and K. Caine, "Users' Privacy Concerns About Wearables," in *Financial Cryptography and Data Security*, M. Brenner, N. Christin, B. Johnson, and K. Rohloff, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 231–244.

[11] T. Denning, Z. Dehlawi, and T. Kohno, "In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-mediating Technologies," in *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: ACM, 2014, pp. 2377–2386. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557352

[12] Y. Rashidi, T. Ahmed, F. Patel, E. Fath, A. Kapadia, C. Nippert-Eng, and N. M. Su, ""You don't want to be the next meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography," in *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. Baltimore, MD: USENIX Association, 2018, pp. 143–157. [Online]. Available: https://www.usenix.org/conference/soups2018/presentation/rashidi

[13] J. M. Such, J. Porter, S. Preibusch, and A. Joinson, "Photo Privacy Conflicts in Social Media: A Large-scale Empirical Study," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 3821–3832. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025668

[14] T. Orekondy, B. Schiele, and M. Fritz, "Towards a visual privacy advisor: Understanding and predicting privacy risks in images," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[15] V. Garg, S. Patil, A. Kapadia, and L. J. Camp, "Peer-produced privacy protection," in *IEEE International Symposium on Technology and Society (ISTAS)*, Jun. 2013, pp. 147–154.

[16] Y. Rashidi, A. Kapadia, C. Nippert-Eng, and N. M. Su, ""It's easier than causing confrontation": Sanctioning Strategies to Maintain Social Norms of Content Sharing and Privacy on Social Media," *To appear in the Proceedings of the ACM Journal: Human-Computer Interaction: Computer Supported Cooperative Work and Social Computing (CSCW '20)*, 2020.

[17] Y. Pu and J. Grossklags, "Using conjoint analysis to investigate the value of interdependent privacy in social app adoption scenarios," *Proceedings of the International Conference on Information Systems (ICIS 2015)*, 2015.

[18] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair, "Over-exposed?: Privacy Patterns and Considerations in Online and Mobile Photo Sharing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 357–366. [Online]. Available: http://doi.acm.org/10.1145/1240624.1240683

[19] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia, "Privacy Behaviors of Lifeloggers Using Wearable Cameras," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. New York, NY, USA: ACM, 2014, pp. 571–582. [Online]. Available: http://doi.acm.org/10.1145/2632048.2632079

[20] A. Efrati, "Read Congress's Letter About Google Glass Privacy," 2013. [Online]. Available: https://blogs.wsj.com/digits/2013/05/16/congress-asks-google-about-glass-privacy/

[21] Office of the Privacy Commissioner of Canada, "Data protection authorities urge Google to address Google Glass concerns," 2013. [Online]. Available: https://www.priv.gc.ca/en/opc-news/news-and-announcements/2013/nr-c_130618/

[22] L. P. Tosun, "Motives for Facebook use and expressing true self on the Internet," *Computers in Human Behavior*, vol. 28, no. 4, pp. 1510–1517, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563212000842

[23] S. Han, J. Min, and H. Lee, "Antecedents of social presence and gratification of social connection needs in SNS: A study of Twitter users and their mobile and non-mobile usage," *International Journal of Information Management*, vol. 35, no. 4, pp. 459–471, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401215000407

[24] Y.-C. Ku, R. Chen, and H. Zhang, "Why do users continue using social networking sites? An exploratory study of members in the United States and Taiwan," *Information & Management*, vol. 50, no. 7, pp. 571–581, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378720613000839

[25] Y. Zhang, L. S.-T. Tang, and L. Leung, "Gratifications, Collective Self-Esteem, Online Emotional Openness, and Traitlike Communication Apprehension as Predictors of Facebook Uses," *Cyberpsychology, Behavior, and Social Networking*, vol. 14, no. 12, pp. 733–739, 2011. [Online]. Available: https://doi.org/10.1089/cyber.2010.0042

[26] Google Street View, "Image acceptance and privacy policies," 2018, retrieved March 07, 2018 from https://www.google.com/streetview/privacy/.

[27] R. Hasan, E. Hassan, Y. Li, K. Caine, D. J. Crandall, R. Hoyle, and A. Kapadia, "Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: ACM, 2018, pp. 47:1–47:13. [Online]. Available: http://doi.acm.org/10.1145/3173574.3173621

[28] T. Orekondy, M. Fritz, and B. Schiele, "Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2018.

[29] P. Aditya, R. Sen, P. Druschel, S. Joon Oh, R. Benenson, M. Fritz, B. Schiele, B. Bhattacharjee, and T. T. Wu, "I-Pic: A Platform for Privacy-Compliant Image Capture," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '16. New York, NY, USA: ACM, 2016, pp. 235–248. [Online]. Available: http://doi.acm.org/10.1145/2906388.2906412

[30] L. Zhang, K. Liu, X.-Y. Li, C. Liu, X. Ding, and Y. Liu, "Privacy-friendly Photo Capturing and Sharing System," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '16. New York, NY, USA: ACM, 2016, pp. 524–534. [Online]. Available: http://doi.acm.org/10.1145/2971648.2971662

[31] M. Ra, S. Lee, E. Miluzzo, and E. Zavesky, "Do Not Capture: Automated Obscurity for Pervasive Imaging," *IEEE Internet Computing*, vol. 21, no. 3, pp. 82–87, 5 2017.

[32] J. Shu, R. Zheng, and P. Hui, "Cardea: Context-Aware Visual Privacy Protection from Pervasive Cameras," *arXiv preprint arXiv:1610.00889*, 2016. [Online]. Available: http://arxiv.org/abs/1610.00889

[33] ——, "Your Privacy Is in Your Hand: Interactive Visual Privacy Control with Tags and Gestures," in *Communication Systems and Networks*, N. Sastry and S. Chakraborty, Eds. Cham: Springer International Publishing, 2017, pp. 24–43.

[34] C. Bo, G. Shen, J. Liu, X.-Y. Li, Y. Zhang, and F. Zhao, "Privacy.Tag: Privacy Concern Expressed and Respected," in *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '14. New York, NY, USA: ACM, 2014, pp. 163–176. [Online]. Available: http://doi.acm.org/10.1145/2668332.2668339

[35] S. Li, W. Deng, and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *2017*

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 2584–2593.
[36] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," *arXiv preprint arXiv:1812.08008*, 2018.
[37] A. J. Perez, S. Zeadally, and S. Griffith, "Bystanders' Privacy," *IT Professional*, vol. 19, no. 3, pp. 61–65, 2017.
[38] V. Tiscareno, K. Johnson, and C. Lawrence, "Systems and Methods for Receiving Infrared Data with a Camera Designed to Detect Images based on Visible Light," 2011. [Online]. Available: http://www.google.com/patents/US20110128384
[39] K. N. Truong, S. N. Patel, J. W. Summet, and G. D. Abowd, "Preventing Camera Recording by Designing a Capture-Resistant Environment," in *UbiComp 2005: Ubiquitous Computing*, M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 73–86.
[40] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling, "PrivacEye: Privacy-Preserving First-Person Vision Using Image Features and Eye Movement Analysis," *arXiv preprint arXiv:1801.04457*, 2018.
[41] A. Perez, S. Zeadally, L. Matos Garcia, J. Mouloud, and S. Griffith, "FacePET: Enhancing Bystanders Facial Privacy with Smart Wearables/Internet of Things," *Electronics*, vol. 7, no. 12, p. 379, 2018.
[42] M. Dimiccoli, J. Marín, and E. Thomaz, "Mitigating Bystander Privacy Concerns in Egocentric Activity Recognition with Deep Learning and Intentional Image Degradation," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–18, 1 2018. [Online]. Available: https://doi.org/10.1145/3161190
[43] A. Li, Q. Li, and W. Gao, "PrivacyCamera: Cooperative Privacy-Aware Photographing with Mobile Phones," in *2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 6 2016, pp. 1–9.
[44] B. Henne, C. Szongott, and M. Smith, "SnapMe if You Can: Privacy Threats of Other Peoples' Geo-tagged Media and What We Can Do About It," in *Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, ser. WiSec '13. New York, NY, USA: ACM, 2013, pp. 95–106. [Online]. Available: http://doi.acm.org/10.1145/2462096.2462113
[45] F. Li, Z. Sun, A. Li, B. Niu, H. Li, and G. Cao, "HideMe: Privacy-Preserving Photo Sharing on Social Networks," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 4 2019, pp. 154–162.
[46] J. He, B. Liu, D. Kong, X. Bao, N. Wang, H. Jin, and G. Kesidis, "PuPPIeS: Transformation-Supported Personalized Privacy Preserving Partial Image Sharing," in *IEEE International Conference on Dependable Systems and Networks*. Atlanta, Georgia USA: IEEE Computer Society, 2014.
[47] M.-R. Ra, R. Govindan, and A. Ortega, "P3: Toward Privacy-preserving Photo Sharing," in *USENIX Conference on Networked Systems Design and Implementation*, ser. nsdi'13. Berkeley, CA, USA: USENIX Association, 2013, pp. 515–528.
[48] Q. Sun, L. Ma, S. Joon Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and Effective Obfuscation by Head Inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2018.
[49] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine, "Effectiveness and Users' Experience of Obfuscation as a Privacy-Enhancing Technology for Sharing Photos," *Proceedings of the ACM: Human Computer Interaction (PACM)*, 2018.
[50] R. Hasan, Y. Li, E. Hassan, K. Caine, D. J. Crandall, R. Hoyle, and A. Kapadia, "Can privacy be satisfying? On improving viewer satisfaction for privacy-enhanced photos using aesthetic transforms," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, vol. 14. ACM, 2019, p. 25. [Online]. Available: http://doi.acm.org/10.1145/3290605.3300597
[51] E. T. Hassan, R. Hasan, P. Shaffer, D. Crandall, and A. Kapadia, "Cartooning for Enhanced Privacy in Lifelogging and Streaming Videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 7 2017, pp. 1333–1342.
[52] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The Open Images Dataset V4," *International Journal of Computer Vision*, 2020. [Online]. Available: https://doi.org/10.1007/s11263-020-01316-z
[53] "Qualtrics." [Online]. Available: https://www.qualtrics.com
[54] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011. [Online]. Available: https://doi.org/10.1177/1745691610393980
[55] R. M. Khan and M. A. Khan, "Academic sojourners, culture shock and intercultural adaptation: A trend analysis," *Studies About Languages*, vol. 10, pp. 38–46, 2007.
[56] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data." *Psychological methods*, vol. 17, no. 3, pp. 437–455, 9 2012.
[57] D. Liu, R. G. Bias, M. Lease, and R. Kuipers, "Crowdsourcing for usability testing," *Proceedings of the American Society for Information Science and Technology*, vol. 49, no. 1, pp. 1–10, 2012. [Online]. Available: http://dx.doi.org/10.1002/meet.14504901100
[58] A. Field, J. Miles, and Z. Field, *Discovering statistics using R*. Sage publications, 2012.
[59] A. G. Yong and S. Pearce, "A beginner's guide to factor analysis: Focusing on exploratory factor analysis," *Tutorials in quantitative methods for psychology*, vol. 9, no. 2, pp. 79–94, 2013.
[60] "12.6 - Reducing Structural Multicollinearity." [Online]. Available: https://newonlinecourses.science.psu.edu/stat501/node/349/
[61] J. W. Osborne, A. B. Costello, and J. T. Kellow, "Best practices in exploratory factor analysis," *Best practices in quantitative methods*, pp. 86–99, 2008.
[62] L. Qu, G. Ferraro, L. Zhou, W. Hou, and T. Baldwin, "Named entity recognition for novel types by transfer learning," *arXiv preprint arXiv:1610.09914*, 2016.
[63] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244*, 2016.
[64] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2014.
[65] G. Gkioxari, R. Girshick, and J. Malik, "Contextual Action Recognition With R*CNN," in *The IEEE International Conference on Computer Vision (ICCV)*, 12 2015.
[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[68] P. Hu and D. Ramanan, "Finding Tiny Faces," *CoRR*, vol. abs/1612.0, 2016. [Online]. Available: http://arxiv.org/abs/1612.04402
[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.
[70] J. Cohen, "Statistical power analysis for the social sciences," 1988.
[71] D. Anthony, C. Campos-Castillo, and C. Horne, "Toward a sociology of privacy," *Annual Review of Sociology*, vol. 43, no. 1, pp. 249–269, 2017. [Online]. Available: https://doi.org/10.1146/annurev-soc-060116-053643
[72] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2017, vol. 39, no. 6, pp. 1137–1149.

APPENDIX

A. Predictive power of each feature

In section V-B, we saw that the features are associated with the classification rationales (Table III and Table IV). Next, we want to investigate how effectively the features can distinguish between *subject* and *bystander*. Results of logistic regression analyses using each of the features individually as predictors are reported in Table VII. The χ^2 statistic indicates how well the data fit the model, where higher values indicate better fit. The value of the R^2 statistic refers to the amount of variance of the outcome variable that was explained by the predictor

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
EFFECTIVENESS OF VISUAL FEATURES USED INDIVIDUALLY AS PREDICTORS TO CLASSIFY *subject* AND *bystander*. ALL χ^2 STATISTICS ARE SIGNIFICANT AT $p < 0.0001$ LEVEL.

Predictor	Odds ratio	[2.5%	97.5%]	χ^2	R^2
Replaceable	0.09	0.07	0.10	2254.41	0.44
Awareness	5.19	4.66	5.78	1476.37	0.29
Willingness	4.38	3.96	4.86	1247.30	0.24
Pose	4.48	4.01	5.00	1146.42	0.22
Comfort	4.05	3.66	4.48	1121.78	0.22
Size	5.23	4.52	6.05	960.15	0.19
Distance	0.31	0.29	0.34	930.95	0.18
Number of people	0.50	0.46	0.54	410.43	0.08
Photographer intention	0.53	0.49	0.57	330.39	0.06
Photo place	1.41	1.32	1.51	101.60	0.02

variable. Note that *Replaceable* has the largest values for both of the statistics, which is intuitive since it is almost a synonym for *being a bystander*. For each predictor, the *Odds ratio* with 95% confidence interval is also presented in Table VII. *Odds ratio* refers to the effect of increasing a predictor's variable by one unit to the outcome variable in a multiplicative scale. For example, increasing the value for *Pose* by one unit will *increase* the odds of a person of being classified as a *subject* by 4.48 times than before. On the other hand, increasing the value for *Replaceable* by one unit will *decrease* the odds of a person of being classified as a *subject* by 11.11 times than before. When used as individual predictors, the features *Replaceable*, *Awareness*, *Willingness*, *Pose*, and *Comfort* all have reasonably high effects on the outcome variable and the data fit the model well enough. But *Photo place* is not a very effective predictor (OR=1.41, χ^2=101.6). The *Size* feature has large effect on the outcome, but using this as an individual predictor it may be noisy as suggested by the lower χ^2 value.

B. Correlation among pairs of features

Table VIII shows Pearson's product moment correlation coefficients (r) between pairs of features. Almost all pairs of features have medium to high correlations between them [70]. In particular, *Awareness* is highly correlated with most of the other features, suggesting that they collectively contain the same information as the 'Awareness' feature.

Table IX shows the VIF for each feature before and after removing the highly correlated 'Awareness' feature.

C. Predicting high-level concepts from the proxy features

As detailed in the Section IV-C3, we infer the *high-level concepts* using the proxy features – human related features, body-pose features, and emotion – using linear regression models. For each of the *high-level concepts*, the mean and standard deviations for training loss, *mean squared error (MSE)*, and *mean absolute error (MAE)* across a 10-fold cross-validation of the regression models are shown in Table X. The error values are interpreted in relation to the range of scores of the outcome variable, since the same error score would indicate a good or bad model depending on whether the range is large or small, respectively. In our case, all the concepts except *Willingness* have the same range of possible values (-3 to 3), and so the prediction errors for them can be compared. *Photographer's intention* has the highest loss and prediction errors. This was expected given that it is more nuanced than the other concepts, and highly depends on the overall context of the image and interactions among people in it. Since we only used features from the cropped portion of the image containing a single person for prediction, the loss and errors go higher. On average *Comfort* could be predicted with the highest accuracy. All the other concepts have about the same losses and prediction errors. Finally, *Willingness* has

TABLE VIII
CORRELATION COEFFICIENTS BETWEEN PAIRS OF VISUAL FEATURES. EACH COEFFICIENT IS SIGNIFICANT AT $p < .001$ LEVEL.

Feature1	Feature2	Correlation coefficient (r)
Awareness	Pose	0.88
	Comfort	0.75
	Willingness	0.79
	Replacable	-0.57
	Size	0.45
	Distance	-0.37
Pose	Comfort	0.73
	Willingness	0.76
	Replacable	-0.48
	Size	0.42
	Distance	-0.34
Comfort	Willingness	0.86
	Replacable	-0.49
	Size	0.37
	Distance	-0.32
Willingness	Replacable	-0.52
	Size	0.39
	Distance	-0.33
Replacable	Size	-0.44
	Distance	0.42
	Number of people	0.31
Size	Distance	-0.48
	Number of people	-0.43

TABLE IX
VARIANCE INFLATION FACTOR (VIF) OF PREDICTOR VARIABLES WHEN ALL PREDICTORS WERE USED (INITIAL VIF) AND AFTER *Awareness* WAS REMOVED (UPDATED VIF).

Variable	Initial VIF	Updated VIF
Awareness	5.80	-
Pose	4.67	2.62
Comfort	4.24	4.23
Willingness	5.01	4.72
Photographer intention	1.11	1.1
Replaceable	1.77	1.73
Photo place	1.14	1.13
Size	1.71	1.7
Distance	1.42	1.42
Number of people	1.27	1.27

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.

TABLE X
RESULTS OF PREDICTING *high-level concepts* USING IMAGE DATA. COLUMNS SHOW MEANS AND STANDARD DEVIATIONS OF *loss*, *mean absolute error (MAE)*, AND *mean squared error (MSE)* OF A 10-FOLD CROSS-VALIDATION.

Outcome	Loss		MAE		MSE	
	Mean	SD	Mean	SD	Mean	SD
Awareness	1.79	0.07	1.04	0.02	1.65	0.06
Photographer's intention	2.65	0.15	1.30	0.04	2.47	0.15
Replaceable	1.60	0.08	0.98	0.03	1.46	0.07
Pose	1.99	0.14	1.08	0.05	1.81	0.14
Comfort	0.81	0.05	0.67	0.03	0.72	0.05
Willingness	0.45	0.02	0.50	0.02	0.40	0.02

TABLE XI
PERCENTAGE OF PARTICIPANTS AGREED WITH THE FINAL CLASSIFICATION LABEL AND NUMBER OF PHOTOS WITH THAT AGREEMENT VALUES.

Agreement	Number of photos
33%	256
50%	208
67%	1308
75%	300
100%	1309

a smaller range of possible values (-2 to 2), and accordingly, smaller loss and error values.

D. Agreement among the annotators

Table XI presents the percentages of agreement among the study participants and the number of images for each percentage. We included percentages for which the number of photos are greater than 100.

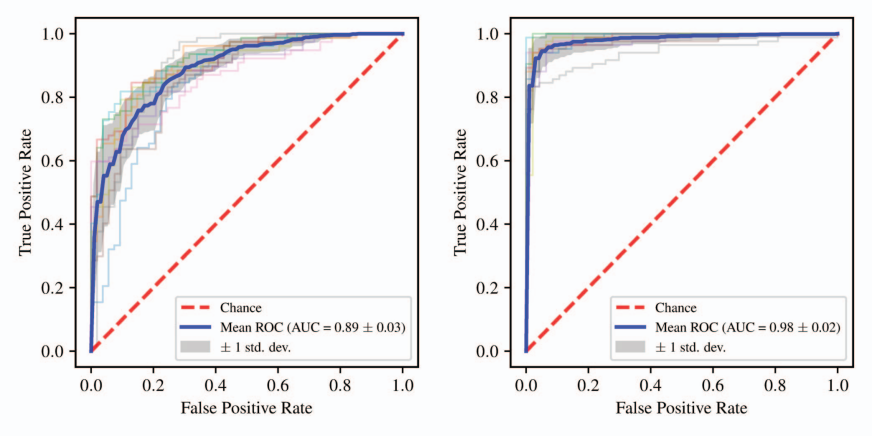
E. Comparing with human annotators

Figure 6 shows Receiver Operating Characteristic (ROC) plots for classifiers trained and tested on images with 67% and 100% agreements among the survey participants.

F. Attention check questions

The two images shown in Fig. 7 were used for attention check questions. We asked **Which of the following statements is true for the person inside the green rectangle in the photo?** with answer options i) There is a person with some of the major body parts visible (such as face, head, torso); ii) There is a person but with no major body part visible (e.g., only hands or feet are visible); iii) There is just a depiction/representation of a person but not a real person (e.g., a poster/photo/sculpture of a person); iv) There is something else inside the box; and v) I don't see any box. Since the persons in the bounding boxes are clearly visible, if any survey participant responded with any option other than the first one, we marked it as wrong.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.



(a) 67% agreement (b) 100% agreement

Fig. 6. Receiver operating characteristic (ROC) plots for classifiers trained and tested on images with (a) 67% agreement and (b) 100% agreement among the survey participants.

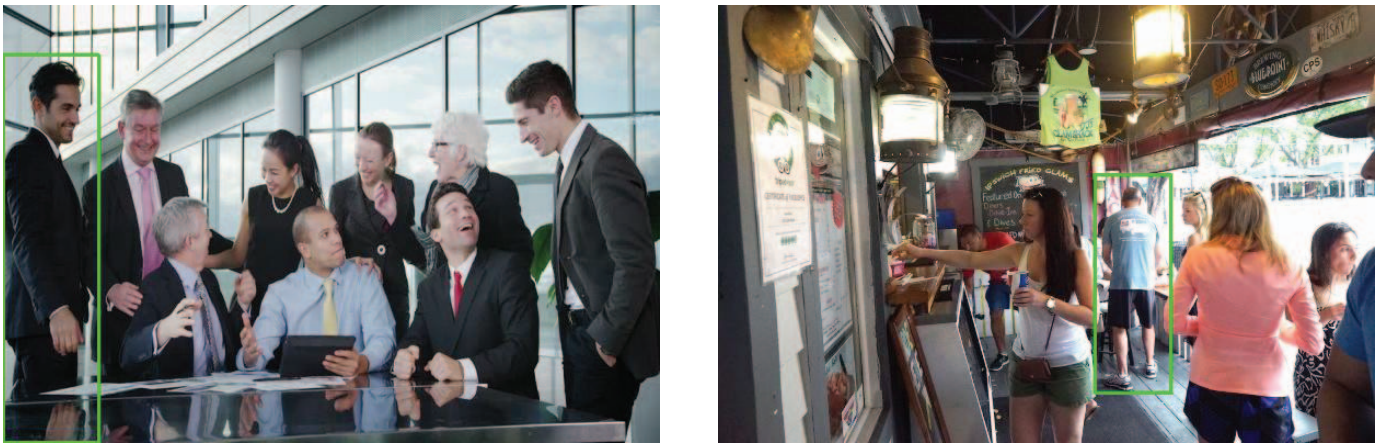


Fig. 7. Images used for attention check questions.

Authorized licensed use limited to: IEEE Xplore. Downloaded on September 26,2022 at 06:04:53 UTC from IEEE Xplore. Restrictions apply.