Human Attributes Prediction Under Privacy-preserving Conditions

Anshu Singh Shaojing Fan Mohan Kankanhalli
School of Computing, National University of Singapore
{anshu|fansj|mohan}@comp.nus.edu.sg

Figure 1: An overview of our framework. (a) We first constructed the Diversity in People and Context Dataset (DPaC) of 8,072 images with human annotations for 10,390 targets. (b) We then conducted psychophysics human studies with eye-tracking to understand the role of context in human attributes prediction under different de-identification privacy scenarios. (c) We designed a deep learning model (CHAPNet) guided by our psychophysics findings for human attributes prediction under different privacy settings. (d) Finally, we investigated the privacy-utility trade-off by visualizing our model.

ABSTRACT

Human attributes prediction in visual media is a well-researched topic with a major focus on human faces. However, face images are often of high privacy concern as they can reveal an individual's identity. How to balance this trade-off between privacy and utility is a key problem among researchers and practitioners. In this study, we make one of the first attempts to investigate the human attributes (emotion, age, and gender) prediction under the different de-identification (eyes, lower-face, face, and head obfuscation) privacy scenarios. We first constructed the Diversity in People and Context Dataset (DPaC). We then performed a human study with eye-tracking on how humans recognize facial attributes without the presence of face and context. Results show that in an image, situational context is informative of a target's attributes. Motivated by our human study, we proposed a multi-tasking deep learning model—Context-Guided Human Attributes Prediction (CHAPNet), for human attributes prediction under privacy-preserving conditions. Extensive experiments on DPaC and three commonly used benchmark datasets demonstrate the superiority of CHAPNet in leveraging the situational context for a better interpretation of a target's attributes without the full presence of the target's face.

This work is licensed under a Creative Commons Attribution-ShareAlike International 4.0 License.

MM '21, October 20–24, 2021, Virtual Event, China.

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8651-7/21/10.

<https://doi.org/10.1145/3474085.3475687>

Our research demonstrates the feasibility of visual analytics under de-identification for privacy.

CCS CONCEPTS

• Security and privacy → Privacy protections; • Computing methodologies → Computer vision.

KEYWORDS

Privacy protection, visual media, human attributes prediction, eye-tracking, deep learning

ACM Reference Format:

Anshu Singh Shaojing Fan Mohan Kankanhalli. 2021. Human Attributes Prediction Under Privacy-preserving Conditions. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3474085.3475687>

1 INTRODUCTION

Researchers have been striving to perform visual analytics under privacy-preserving conditions [11, 46, 51]. For example, crowd counting without getting the people model [46], or recognizing human actions under extremely low resolutions so that the identity of the people in the video will not be revealed [14, 40]. The biggest challenge for the above tasks is how to balance the trade-off between task performance and privacy budget. We seek to explore this less-researched but important problem, by making one of the first

attempts to predict the human attributes¹ under the different de-identification (eyes, lower-face, face, and head obfuscation) privacy scenarios. We introduce the concept of *partial privacy*: obfuscating faces in a subset of training images, and demonstrate that it can give better performance than *strict privacy*: obfuscating faces in all training images.

Our research is based on prior works in multiple disciplines. Psychologists have underscored the significance of situational context² in emotion recognition [2, 12, 20, 24, 33]. Yang et al. [48], demonstrated that a pre-trained model on original and blurred faces in the ImageNet [16] has minimal impact on the classification accuracy of face-centric tasks such as human attributes classification. Inspired by previous research, we first constructed the Diversity in People and Context dataset (DPaC), a new dataset of 8,072 RGB images, with extensive *perceived*³ human annotations (we have 5 age categories: *Child*, *Adolescent*, *Young adult*, *Middle-aged adult* and *Older adult*), binary gender, 7 emotion categories: *Happiness*, *Anger*, *Surprise*, *Disgust*, *Fear*, *Sadness* and *Neutral*, 3 skin color categories: *Light*, *Medium* and *Dark*, 3 personality traits' rating: *Friendliness*, *Trustworthiness*, and *Calmness*, and 10 social relation categories) for 10,390 targets with low face blur value. We then conducted an eye-tracking-based human study to explore the subjects' perception of emotion, age, and gender of targets under various privacy settings (e.g., obfuscation of a target's face or surroundings). Results show that situational context can aid in the human estimation of emotion, age, and gender when facial features are not present. Motivated by our human study, we propose the Context-guided Human Attributes Prediction (CHAPNet), a multi-tasking deep learning model which leverages situational context for human attributes prediction under privacy-preserving conditions. Experiments on four datasets demonstrate the advantage of the proposed model in learning the contextual information for human attributes prediction when facial features are removed.

Our work advances the existing research for human attributes prediction, which relies mainly on facial information. Our experimental results in various privacy settings can help measure the privacy-utility trade-off for related computer vision tasks. Moreover, our study encourages computer vision scientists to consider privacy protection in both visual datasets and visual tasks. Our main contributions follow:

- We propose a novel dataset DPaC which 1) provides extensive human attributes annotations (age, gender, emotion, skin color, personality traits, and social relations); and 2) proposes to address the geo-diversity problem⁴ in human attribute-based tasks. With intense labeling and a more balanced geo-diversity, our dataset can help the community with insightful analyses of human behavior, as well as multiple human-related visual tasks, such as social interactions.
- We perform a series of human studies with eye-tracking to explore the subjects' understanding of emotion, age, and

gender estimation of targets under different obfuscation scenarios. Results indicate the significant contribution of situational context in human attributes perception when a face is not present.

- We propose a multi-tasking deep learning model CHAPNet inspired by our human study. It is novel in its approach for human attributes prediction by relying on situational context for a better interpretation of human attributes, especially under privacy-preserving conditions (i.e., trained on all or a fraction of face obfuscated images).
- We are one of the first to investigate the impact of face obfuscation scenarios (eye, lower-face, face, and head) on automated human attributes prediction. We show that the increase in face obfuscation area lowers the model's performance. Such trend is most significant on emotion prediction, compared with age and gender.

The code and dataset with images containing obfuscated faces are available online⁵. The dataset with images of intact faces can be provided upon request.

2 RELATED WORK

2.1 Privacy-preserving Visual Analytics

With pervasive cameras for surveillance or smart home devices, privacy-preserving visual analytics has gradually become very important in recent years [11, 46, 51]. The work in [11] proposed a crowd monitoring method without the need for people models or tracking. Wu and colleagues [46] proposed a privacy-preserving visual recognition method in a unique adversarial training framework. Other studies [14, 40] focus on human action recognition in videos of low resolutions so that video surveillance can be done without compromising targets' identity. A notable challenge in the above tasks is the trade-off between utility and privacy. In our work, we make one of the first attempts for human attributes prediction under privacy-preserving conditions. Our methodology is different from previous studies: we motivate our model from human psychophysics studies. We analyze where people focus during human attributes prediction under face obfuscation and design a model to mimic such human behavior.

2.2 Datasets for Human Attributes Prediction

Most of the existing datasets with human attributes annotations are face-centered without context [23, 25, 26, 29, 52]. They are not suitable for our study, as we focus on images of non-face-centric people in unconstrained environments. The most relevant datasets we found include CAER-S dataset [28], Images of Groups (IoG) dataset [18], and EMOTIC [27] dataset, which are used later in our computational experiments. Among which the EMOTIC dataset is the closest, in our knowledge, with annotated human attributes of targets in unconstrained environments. However, it consists of more than 25% targets in the images under occlusion, which makes the labeling noisy. Furthermore, the emotion, age, and gender labels in EMOTIC are highly imbalanced. Indeed, a majority of current datasets suffer from the over-representation of some groups, like the dominance of Caucasian faces [42, 47]. We attempt to address

¹We refer to human attributes as emotion, age, and gender in this paper.

²We refer to context in an image as everything except for a target's head. Situational context is defined as everything except for a target's body.

³In our paper, a perceived attribute is estimated based on the perception of an observer, rather than its real value.

⁴Geo-diversity problem is the dominance of people in visual data belonging to limited geographical locations [42]

⁵https://github.com/anshu0612/human_attributes_prediction_in_privacy_conditions

the aforementioned limitations in our proposed DPAC dataset. We hope this dataset can be a benchmark for human attributes prediction under privacy-preserving conditions in an unconstrained environment.

2.3 Role of Context in Human attributes Prediction

Psychophysics studies: Psychologists have underscored the significance of situational context in emotion recognition [2, 12, 20, 24, 33]. Gallagher et al. [18], suggested that the presence of a person in a scene positioned in an image is a part of a context. Chen and Whitney [12] designed an experiment to demonstrate that situational context is often sufficient in silent videos to predict a person’s emotion without the presence of the body of a target. However, as Martinez [12] has pointed out, they failed to include knowledge about what particular image cues our visual system uses to infer a target’s emotion. We fill this knowledge gap through our eye-tracking-based experiment by examining the participants’ eye-gaze fixations and saccades. Furthermore, different from the above studies, our study is not limited to emotion prediction; we also expand our research to examine the contribution of context to age and gender perception.

Computational modelling: Recently, there have been some studies using deep learning networks for human emotion recognition by including the extracted features from the context [19, 27, 28, 35, 50]. However, most of the existing methods rely on and encode full facial features, whereas we train and demonstrate our model’s performance without explicitly localizing on face information. The work in [18] argues that by utilizing situational context, age and gender prediction accuracy can be increased. However, their research focuses on images of people in social gatherings, whereas our model works on people in unconstrained environments.

3 THE DIVERSITY IN PEOPLE AND CONTEXT DATASET (DPAC)

Dataset overview: We constructed the Diversity in People and Context dataset (DPAC), a new dataset of 8,072 RGB images, with human annotations for 10,390 targets with low face blur value (i.e., values <0.75 - predicted by Microsoft Face Detect API⁶). DPAC is unique in following ways: 1) It is diverse in various aspects - it consists of images of geo-diverse people in different situations, diverse backgrounds, multiple actions, and varying demographics. 2) It is the first dataset of its kind to address the lack of geo-diversity problems in a dataset [42]. Such a dataset is suitable for our task to study human attributes in realistic settings. With the low face-blurriness of the targets, the impact of face obfuscation can be reliably studied. 3) It has extensive human annotations including perceived emotion, age, and gender labelings, as well as skin color, personality traits, and social relations of targets in images. It provides both face and body bounding boxes. With such intense labelings, further research in this direction can help the community with insightful analyses of human behavior, as well as human social interactions.

⁶<https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-detection>

Human attributes collection and statistics: We crowd-sourced the attributes annotations on online crowd-sourcing (Amazon Mechanical Turk (AMT)) [8] by 601 AMT workers. For quality control, a set of pre-annotated gold standard questions were used for measuring the quality of workers. For any specific category, we used the majority rule to collect the ground truth. In case of no consensus, we got the images re-annotated by 2 more workers; if yet consensus was not reached, we discarded those images. Fig. 2 illustrates the distribution of emotion, age, and gender in DPAC. While ethical and bias considerations related to human attributes are beyond the paper’s scope, as a step towards tackling the growing concern over biases in the dataset [7, 9, 13, 42, 47], we report the distribution over the skin color.

The dataset comes with the train and test split of 80% and 20%, respectively. Readers can refer to Sec. A of the supplementary material for more details.

Figure 2: Targets count for each human attribute in DPAC dataset.

4 HUMAN PSYCHOPHYSICS EXPERIMENTS

Based on previous literature [6, 31, 36, 37, 43], we hypothesize humans have an ability to recognize human attributes with considerable reliability under the absence of facial information. We investigate this hypothesis with the help of our human study with eye-tracking, focusing on understanding the role of situational context in human attributes prediction under privacy-preserving conditions. Refer to Sec. B of the supplementary material for further details.

Setup: We randomly selected 120 images from the DPAC dataset. For each image we created 4 different versions (see Fig. 3.a) – *Cropped Obfuscated*: cropped image leaving out a target’s body and obfuscated face, *All Obfuscated*: all faces were obfuscated, *Target Obfuscated*: only a target’s face was obfuscated, and *No Obfuscation*: every face was visible in an image. An image version was presented in this similar order of divulged image content. We highlighted a target in a red rectangle and used “box gray-in” for face obfuscation. The experiment consisted of three sub-experiments - containing non-overlapping stimuli to tackle the memorability issue [10] - to predict the emotion, age, and gender respectively of a target person. 24 participants from the National University of Singapore observed the series of stimuli presented to them while we recorded their eye movements.

Figure 3: (a) For each stimulus in our human study, we design four different versions with various levels of context information - 1. *Cropped Obfuscated*: cropped image leaving out a target's body and obfuscated face, 2. *All Obfuscated*: all faces were obfuscated, 3. *Target Obfuscated*: only a target's face was obfuscated, 4. *No Obfuscation*: every face was visible in an image. (b) Participants mean accuracy on different versions of stimuli. Error bars represent the standard error of means. The asterisk denotes $*p < 0.05$, $**p < 0.01$.

Analysis: In this subsection, we visualize the situational cues used for attributes prediction with the help of collected eye-tracking data. We report the analysis of variance (ANOVA) with repeated measurements having accuracy as a dependent variable and the different versions of the images excluding the *No Obfuscation* version as the independent variable. *No Obfuscation* version is for obtaining the ground truth for analyses by majority voting.

Observation 1: Situational cues can help in better understanding of age and gender of a target, and its contribution is higher on age than gender.

During a target's gender and age prediction, participants' eye-fixations were majorly weighted towards the bodily cues like visible skin regions, pose, hair and clothing.

Figure 4: Top-row: Heatmaps showing specific situations when participants relied on situational context for age prediction. Bottom-row: Gazemaps showing a few situations when participants relied on situational context for gender prediction. a) Female in the army; b) Limited bodily cues and uncommon hairstyle; c) Children.

As compared to the estimation of age, the situational context has limited impact ($F(2, 60) = 3.75$, $p < 0.05$) for gender estimation. We notice that in some scenarios, participants might have reasoned out from the situational context when bodily cues were misleading, confusing, or insufficient. These scenarios include women wearing military uniforms, gender of children, which can be hard to predict (see Fig. 4).

Observation 2: Objects and background information have significant contribution for inferring a target's emotion.

As shown in Fig. 3, for emotion prediction, there is a significant improvement of 12.5% in the accuracy of the participants from *Cropped Obfuscated* to *All Obfuscated* condition ($F(1, 40) = 10.85$, $p < .005$). Under *All Obfuscated* condition, the information apart from a target, i.e., the background and objects, are visible. The eye saccades jump from bodily cues to the background (See Fig.5 top-row), and vice versa. Moreover, subjects allocate fixations on an object in a target's hand for emotion inference.

Figure 5: Top-row: Visualizations of participants' fixation saccades, suggesting dependency on the background, objects and body for inferring targets' emotions. Bottom-row: Fixations showing contribution of nearby people's emotion for inferring targets' emotions.

Observation 3: The emotions of nearby people play the most significant role in inferring a target's emotion. The emotions of socially interactive persons in proximity to a target have a higher weight in such inference over the emotions of the rest of the people in an image.

As shown in Fig. 3, for emotion prediction, there is a significant improvement of 14.15% in the accuracy of the participants when the faces of other people are made visible from *All Obfuscated* to *Target Obfuscated* condition ($F(1, 40) = 28.61$, $p < .00001$). There is an improvement in the participants' consensus with the eye fixations on the nearby person's face. This elucidates how emotions are socially constructed and are biologically evident [5]. People at the proximity of a target with certain interactions receive a relatively higher intensity of fixation than unrelated people or those far away from a target (See Fig.5 bottom-row). This is consistent with previous research in anthropology [3] and social psychology [21] which stated that comfortable spacing between people depends on social relationship and situation. Furthermore, our observation suggests that another person's simple presence does not elicit more expression of emotion, which is different from previous studies [12, 20]. But it is the emotion carried by a nearby person that makes the difference.

5 CONTEXT GUIDED HUMAN ATTRIBUTES PREDICTION NETWORK

This section proposes the Context-guided Human Attributes Prediction Network (CHAPNet) guided by our human findings. We utilized the psychophysics observations for emotion, age, and gender prediction to design CHAPNet, an end-to-end multi-tasking human attributes classification deep learning model.

Figure 6: (a) Architecture overview of the proposed Context-guided Human Attributes Prediction (CHAPNet) for gender, age, and emotion prediction. The model uses the features extracted from both the situational context branch and target branch for the attributes prediction. (b) Qualitative results with intact face images by CHAPNet trained on only face obfuscated images. In the target branch, B refers to the global features extracted with ResNet-34.

Architecture overview: The architecture of CHAPNet is illustrated in Figure 6.a. It consists of two branches - The target branch and the situational context branch. The target branch extracts features from a cropped target $I_t \in \mathbb{R}^{400 \times 400}$ whereas the situational context branch does feature extraction from an entire image $I_c \in \mathbb{R}^{400 \times 400}$. The two branches' output is merged and passed through three fully connected layers (FC) for gender, age, and emotion prediction, respectively (Figure 6.b). Below we describe the two branches in detail, along with the underpinning observations from our human study.

Situational context branch: Our human study shows that situational context plays a significant role in predicting the emotion, age, and gender of a target person when facial information is removed. We use this information as prior knowledge in designing the situational context branch. Particularly, we break down the features extraction from situational context into target-visual relationship - $f_{relation}$, people and objects - $f_{objects}$, and event - f_{event} representations. The three representations echo the idea from the observations made in our human study that background information like events, scenes, and objects support human attributes' interpretability.

f_{event} : We use the pre-trained ResNet-18 trained on Places dataset [53]. The pre-trained model on such a high-diversity dataset offers a valuable resource to help in event recognition.

$f_{objects}$: For objects and people learning in an image, we use pre-trained ResNet-34 on ImageNet [16].

$f_{relation}$: For effective target-visual relationship learning from the situational context, we use Graph Attention Network (GAT) [44]. A 400×400 input image, I_c is first passed through the adaptive max-pooling for dimensionality reduction. It is then split into patches of dimension 30×30 each, with each patch embedded as a vector by passing through FC layers. These feature vectors form the nodes of the graph. An additional node is added embedding the extracted features from the target's body, I_t . Similar to I_c , I_t is passed through the pooling layer for dimensionality reduction and after passing through FC layers, is represented as a vector. The node representing a target and other nodes representing the different regions of images helps in capturing the visual relationship with nearby people and

objects. Besides, the attention mechanism of GAT helps in learning and allocating weights to the image patches.

Target branch: This stream is inspired by the model in [34], where they designed the pose-guided feature stream for person identification under occlusion. For our purpose, a pose-guided feature stream can help capture information from different regions of a target's body, similar to the participants' eye-saccades imbibing information from a target's body. Furthermore, landmarks with a low confidence score are ignored, preventing feature extraction from the occluded regions of a target. In particular, the extracted global body feature B from backbone ResNet-34 is first fed to pose-guided feature stream. The 18 landmarks generated by OpenPose library help in generating heat maps consisting of a 2D Gaussian centered on the ground truth location. Heat maps of the landmarks with confidence scores lower than 0.2 are set to zero. Each of the 18 heatmaps is then downsampled by bilinear interpolation and then element-wise multiplied by the global feature B , helping in encoding information from the different non-occluded regions of a target. Each feature map is passed through an average pooling layer to produce a feature vector. Finally, max pooling is applied over all feature vectors and then downsampled by average pooling, outputting f_{target} .

For the classification of the attributes, we use the cross-entropy loss for each attribute and then sum the losses. The details on hyper-parameters and other implementation specifications are provided in Sec. C of the supplementary material.

6 HUMAN ATTRIBUTES PREDICTION UNDER PRIVACY-PRESERVING CONDITIONS

6.1 Experiment Settings

Datasets: We curated three datasets with face labels under natural context - EMOTIC [27], CAER-S [28], and Images of Groups (IoG) [18] datasets. CAER-S contains 70k images taken from the video frames of popular TV shows and targets are annotated with 7 discrete emotions. EMOTIC contains 23,751 images with 34,320 targets in unconstrained environments. Each target is annotated

with 26 emotion categories, binary gender, and 3 age categories (children, teenagers, and adults). IoG consists of 5,080 images of groups of people with 28,231 faces labeled with binary gender and 7 age categories.

Face obfuscation scenarios: We consider the 4 commonly used face obfuscation scenarios on social media — eyes, lower face, face, and head — to de-identify people in the image data. We added a lower-face obfuscation scenario due to the pervasive use of masks in the current COVID-19 pandemic situation [30].

We automated the blocking process with the OpenFace library [4], which generates 67 facial landmarks. We report CHAPNet’s performance on all four datasets in Table 2. The purpose is to use these values as a benchmark in the human attributes prediction in the privacy settings.

As shown in Fig. 7, the landmarks are utilized to estimate the 4 regions of a targeted face. We then applied a simple obfuscation technique, “black box-in”, to hide the facial regions. We also included the undetected faces by OpenFace in the training datasets for maintaining consistency in the datasets’ sizes across the privacy scenarios and intact faces that help in comparative evaluations. We consider these undetected faces *safe* against major machine learning models.

Figure 7: Preprocessing of the datasets for creating face obfuscation scenarios. (a) 67 landmarks generated by the OpenFace library [4] for estimating different face regions. (b) A sample of generated face obfuscation scenarios imposed on the detected facial landmarks.

Privacy settings in training images: We study the model performance under the following privacy settings in training images:

- *Target Obfuscation* — similar to the human study Fig.3.a.3, only a target’s face is obfuscated in all training images.
- *All Obfuscation* — similar to the human study 3.a.2, all the faces in the training images are obfuscated. We name this setting as *strict privacy*. This is an ideal setting where every person, including accidental people, in an image is de-identified and a training dataset can reach maximum privacy protection for every person in an image.
- *Partial Obfuscation* — termed *partial privacy*, where only a fraction of training images are face obfuscated.

6.2 Experimental Results

Comparison with state-of-the-art models: When faces are intact (without obfuscation), our model (CHAPNet) gives an accuracy of 88.31% on CAER-S dataset, outperforming all the comparison models [19, 28, 50] by a large margin. CHAPNet is also competitive on the EMOTIC dataset (Table 1, note that the results of the comparison methods are reported by their authors or in other works). For the IoG dataset, unlike ours which is trained on the whole images, the comparison models [1, 15, 17, 22, 32, 38, 39, 41, 49] take cropped

Method	CAER-S	EMOTIC
	accuracy(%)	mAP(%)
Lee et al. [28]	73.51	20.84
Gao et al. [19]	81.31	-
Koski et al. [27]	-	27.38
Zhang et al. [50]	76.73	28.42
Trisha et al. [35]	-	35.48
Ours (CHAPNet)	88.31	28.81

Table 1: Emotion prediction results on benchmark datasets and models. We report accuracy and mean average precision (mAP) on CAER-S and EMOTIC datasets, respectively. The results of the comparison methods are taken from the papers.

and aligned faces as input. Due to the huge difference in training samples distribution, we do not report their results here.

We report CHAPNet’s performance on all 4 datasets in Table 2. The purpose is to use these values as a benchmark in the human attributes prediction in the privacy settings.

Studying the effect of face obfuscation on classification accuracy: We applied each of the 4 face obfuscation scenarios (see subsection 6.1) under target-vs-all conditions i.e., *Target Obfuscated* and *All Obfuscated*. We exhaustively applied each of the privacy scenarios on DPaC, EMOTIC, CAER-S, and IoG datasets. For the CAER-S dataset, since more than 83% of the images (refer to Fig. 8) contain a single face; therefore for completeness we consider the extreme case, i.e., *All Obfuscated*.

Figure 8: Statistics of faces in our DPaC dataset and other benchmark datasets.

As reported in Table 2, with the increase in the face obfuscation area, there is a monotonic decrease in the performance of human attributes in all the datasets. However, in the EMOTIC dataset, the decrease rate is slow compared to the other datasets. This is because 42% of the images contained undetected faces. This suggests that the distribution of a dataset and attributes categories is an essential factor to consider when inferring the accuracy.

Besides, there is a greater decrease rate of classification accuracy on emotion prediction, followed by age and then gender. This aligns with our human study results, where we showed for emotion and age estimation, the face is the most important aspect. For *Target Obfuscated* vs. *All Obfuscated*, we see no common pattern in either of the conditions giving better performance. This indicates the minimal impact on a target’s human attributes classification accuracy when all the faces are obfuscated.

It is interesting to note that for each attribute, even under the best privacy scenario, i.e., head obfuscation, the model can perform far better than the chance level, i.e., 50% accuracy. This is very

Dataset	DPaC						CAER-S	EMOTIC						IoG			
Attribute	Age		Gender		Emotion		Emotion	Age		Gender		Emotion		Age		Gender	
Condition	TO	AO	TO	AO	TO	AO	AO	TO	AO	TO	AO	TO	AO	TO	AO	TO	AO
B_e	63.36	63.11	89.58	88.38	68.66	70.15	85.83	82.80	83.94	72.22	74.39	27.67	27.72	68.72	68.81	80.08	80.95
B_l	60.51	61.37	83.31	84.09	56.27	55.98	83.52	82.70	83.67	72.01	74.24	27.09	27.21	66.44	68.27	79.47	79.37
B_f	57.52	56.02	81.49	81.39	55.64	54.87	82.62	82.55	83.65	70.94	74.02	26.96	26.86	66.29	66.97	78.11	79.29
B_h	51.74	52.41	73.72	77.39	55.01	54.73	77.64	81.56	83.30	70.12	73.36	26.63	26.56	64.23	64.70	76.59	76.83
Intact	66.73		91.12		72.66		88.31	83.43		76.19		28.81		70.97		81.27	

Table 2: CHAPNet’s human attributes classification performance trained with images under the four face obfuscated scenarios. Each scenario is further reported under the *Target Obfuscated* – TO – and *All Obfuscated* – AO – privacy settings. The values are reported with intact faces on the test-set of the respective datasets. Face obfuscation scenarios - B_e : Blocked eyes, B_l : Blocked lower face, B_f : Blocked face, B_h : Blocked head. Except for emotion prediction on the EMOTIC dataset, where the evaluation metric is mAP (%), accuracy (%) is reported for all the other values.

encouraging and hints at a promising future for our study, i.e., accurate attributes prediction without compromising any individual’s privacy.

6.3 Analysis

Does the model throw the hidden facial features under the *strict privacy* setting?

We hope to know if the model under such a condition throws the hidden facial features during testing. We answer this question with the help of the visualizations generated using Score-CAM [45]. We found that the model throws the hidden facial features, but learns contextual information other than facial features. Fig. 9 shows that the model picks features from the visible bodily cues like hairs, skin, clothing, accessories, and so on, to accomplish the task. Such model behavior is relatable to the human behavior we saw in Observation 1 of our human study. Participants relied heavily on bodily cues for age and gender prediction under the face obfuscation scenario. Humans and machine learning models both strive to achieve better performance in their tasks.

Figure 9: Visualizations under the *strict privacy*. With the increase in face obfuscated area, the trained model learns to suppress features from blocked face regions but extracts features from the visible face regions and bodily cues. This is consistent with the observations we made in our human study where participants relied on bodily cues for the human attributes prediction under the face obfuscation.

Can the model perform better under *partial privacy*?

Fig. 9 shows that the trained model on the *strict privacy* setting

%	Age		Gender		Emotion	
100	56.02	Δ	81.39	Δ	54.87	Δ
75	62.73 ± 0.66	7.09	87.43 ± 0.54	6.65	59.64 ± 0.44	4.91
50	64.00 ± 0.41	7.98	88.80 ± 0.98	7.41	66.19 ± 0.11	11.32
25	65.42 ± 0.09	9.40	90.05 ± 0.54	8.66	70.80 ± 1.64	15.93
0	66.73	10.71	91.12	9.73	72.66	17.79

Table 3: Model’s performance accuracy under *partial privacy* and face obfuscation scenario. The results under *partial privacy* is better than that under the *strict privacy*. The *partial privacy* experiments are repeated three times; the reported values are the mean accuracy and standard error. Δ denotes the increase in accuracy with respect to the *strict privacy*.

ignores the blocked facial features which are essential for the classification task. Such behavior preserves one’s protected facial features. However, the drop in performance might not be satisfactory for certain high-stake applications. Thus we propose an alternative, which is to apply *strict privacy* to only a fraction of images in the training dataset, i.e., *partial privacy*. In real-life scenarios, the unobfuscated images can contain people who do not have privacy concerns about their image usage, e.g., donors, consented users, public figures, and so on.

As shown in Table 3, the *partial privacy* generate better performance than *strict privacy*. With 25% of images with *All Obfuscated* setting, the model reports a less than 2% drop in the accuracy of the human attributes for training with 0% privacy, i.e., with intact faces. This suggests the superiority of our model in learning contextual information, especially under privacy-preserving conditions.

There is a considerable decline in emotion prediction performance as the percentage of *partial privacy* increases, more than that in age and gender. This might be due to suppressed facial features with increased obfuscated images fraction in the training data. Our human studies show that emotion and age are significantly impacted by the loss of facial information. With an increased amount of hidden faces, the model loses confidence in facial features thereby suppressing them (refer to Fig. C.1).

In summary, *partial privacy* can be considered over *strict privacy* for a training dataset in privacy settings. This is because our empirical results show that *partial privacy* can give a significant performance boost in the human attributes classification over *strict privacy*.

Targets of all sizes – how does the ratio of target and image obfuscated impact classification accuracy?

Another important thing to understand is the fraction of target

Input Branches and Streams	Age acc (%)	Gender acc (%)	Emotion acc (%)
TB	50.68	76.66	50.19
TB + SCB - O	64.18	89.3	70.97
TB + SCB - E	66.44	91.03	71.61
TB + SCB - V	61.14	88.52	61.72
TB + SCB	66.73	91.66	72.66

Table 4: Ablation study results for emotion, age, and gender prediction. TB: Target branch, SCB: Situational context branch, O: People and objects stream, E: Event stream, V: Target-visual relationship stream.

and image obfuscated — considering *strict privacy* — and its impact on the human attributes classification accuracy. We calculated an obfuscated target by dividing the face bounding box by the body bounding box. Similarly, for image obfuscation, we summed all the detected face bounding boxes and divided them by an image’s dimensions. As shown in Figure 10, we classified the fractions into bins for target and image obfuscation, respectively. Surprisingly, we do not see a drop in the accuracy with the increase of the ratio for gender and age performances in target and image obfuscated cases. However, the increased ratio of occlusion impacts emotion prediction. This suggests that the increased fraction of occlusion impacts emotion more than age and gender. Such difference might be because visible skin on the forehead and hair can still be leveraged for inferring gender and age, but not for emotion.

Figure 10: Accuracy v.s. the obfuscated fraction of a target (left), whole image (right) under *strict privacy* conditions on the DPac test set. n denotes the cardinality of the images.

6.4 Ablation study

We conducted ablation studies on the DPac dataset to analyze the contribution of different branches in our model, and try to investigate why our model is suitable for human attributes prediction under privacy-preserving conditions. The ablation study results are reported in Table 4, which shows that best estimates of emotion, age, and gender can be accomplished by incorporating the situational context with all of the three streams – people and object, event, and target-visual relationship. Moreover, the results underpin the human study’s observations that emphasized the situational context’s role in human attributes prediction.

7 DISCUSSION

With the incremental face obfuscated regions, there is a drop in the classification accuracy. Nevertheless, even under the extreme privacy scenario, i.e., *strict privacy* with full head obfuscation, our

model can achieve accuracy better than the chance level. This is remarkable, given the human attributes prediction significantly relies on facial features. It shows CHAPNet’s efficacy in gauging contextual cues beyond faces to interpret the human attributes. Also, we observed failure scenarios where situational context is incongruent and confusing or a target is under a major occlusion by other subjects (refer to the supplementary material). However, these scenarios can be addressed in future work with possible solutions like reasoning with multimodal learning. Moreover, our model design can be used to analyze other human attributes such as race/ethnicity, social relations, and personality traits in privacy settings.

Our results show a minimal difference in CHAPNet’s performance between *Target obfuscated* and *All obfuscated* conditions, suggesting further its efficacy in learning contextual cues for human attributes prediction. Further, as deep learning models thrive on large-scale datasets, getting people’s consent on a large scale is challenging [7]. Our results show that with 25% *partial privacy*, one can achieve intact-face level performance with less than a 2% accuracy drop in human attributes prediction. Our results also suggest that *strict privacy* might not be suitable for human attributes like emotion since it is greatly impacted by face obfuscations. Overall, our results are encouraging to seriously consider privacy in visual datasets.

8 CONCLUSION

This study addresses an important but challenging problem with potential future applications involving data analytics - human attributes prediction under privacy-preserving conditions. A novel dataset, Diversity in People and Context, with diversity in context and multiple labeled human attributes of people in natural settings is proposed. A series of human studies with eye tracking showed a calculated joint analysis from various contextual cues could be leveraged to infer a person’s emotion, age, and gender. Guided by our human study, we proposed the CHAPNet model for automated human attributes prediction. We then studied the privacy-utility trade-off in the human attributes classification under the various privacy settings. Apart from advocating privacy in the image dataset, we encourage extending our model design and experimental results to video analysis. This is beneficial for data analysis with video surveillance where one can consider an acceptable amount of the general public’s facial privacy and achieve satisfactory performance in their task. Today, artificial intelligence systems have started becoming better at their task; it is high time to consider privacy protection in such tasks.

Acknowledgements We thank Shen Zhiqi for his generous help in need. We are grateful to Departmental Ethics Review Committee at the School of Computing for their aid on the IRB application.

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

REFERENCES

- [1] Nihad A Abdalrady and Saleh Aly. 2020. Fusion of Multiple Simple Convolutional Neural Networks for Gender Classification. In *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*. IEEE, 251–256.
- [2] Lior Abramson, Rotem Petranker, Inbal Marom, and Hillel Aviezer. 2020. Social interaction context shapes emotion recognition through body language, not facial expressions. *Emotion* (2020).
- [3] Irwin Altman. 1975. The environment and social behavior: privacy, personal space, territory, and crowding. (1975).
- [4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.
- [5] Lisa Feldman Barrett. 2012. Emotions are real. *Emotion* 12, 3 (2012), 413.
- [6] Irving Biederman. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review* 94, 2 (1987), 115.
- [7] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1536–1546.
- [8] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? (2016).
- [9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
- [10] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. 2015. Intrinsic and extrinsic effects on image memorability. *Vision research* 116 (2015), 165–178.
- [11] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–7.
- [12] Zhimin Chen and David Whitney. 2019. Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences* 116, 15 (2019), 7559–7564.
- [13] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).
- [14] Ji Dai, Behrouz Saghaei, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. 2015. Towards privacy-preserving recognition of human activities. In *2015 IEEE international conference on image processing (ICIP)*. IEEE, 4238–4242.
- [15] Afshin Dehghan, Enrique G Ortiz, Guang Shu, and Syed Zain Masood. 2017. Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280* (2017).
- [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [17] Yuan Dong, Yanan Liu, and Shiguo Lian. 2016. Automatic age estimation based on deep learning algorithm. *Neurocomputing* 187 (2016), 4–10.
- [18] Andrew C Gallagher and Tsuhan Chen. 2009. Understanding images of groups of people. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 256–263.
- [19] Qinqian Gao, Hanxin Zeng, Gen Li, and Tong Tong. 2021. Graph Reasoning-Based Emotion Recognition Network. *IEEE Access* 9 (2021), 6488–6497.
- [20] Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. 2018. Context is everything (in emotion research). *Social and Personality Psychology Compass* 12, 6 (2018), e12393.
- [21] Edward T Hall. 1963. A system for the notation of proxemic behavior 1. *American anthropologist* 65, 5 (1963), 1003–1026.
- [22] Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916* (2016).
- [23] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*.
- [24] Jacob Israelashvili, Ran R Hassin, and Hillel Aviezer. 2019. When emotions run high: A critical role for context in the unfolding of dynamic, real-life facial affect. *Emotion* 19, 3 (2019), 558.
- [25] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.
- [26] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.
- [27] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence* 42, 11 (2019), 2755–2766.
- [28] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10143–10152.
- [29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebrities attributes (celeba) dataset. *Retrieved August 15, 2018* (2018), 11.
- [30] Mohamed Loey, Gunasekaran Manogaran, Mohamed Hamed N Taha, and Nour Eldeen M Khalifa. 2021. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement* 167 (2021), 108288.
- [31] Nikos K Logothetis and David L Sheinberg. 1996. Visual object recognition. *Annual review of neuroscience* 19, 1 (1996), 577–621.
- [32] Jordi Mansanet, Alberto Albiol, and Roberto Paredes. 2016. Local deep neural networks for gender recognition. *Pattern Recognition Letters* 70 (2016), 80–86.
- [33] Aleix M Martinez. 2019. Context may reveal how you feel. *Proceedings of the National Academy of Sciences* 116, 15 (2019), 7169–7171.
- [34] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 542–551.
- [35] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14234–14243.
- [36] Nicolas Pinto, N Majaj, Youssef Barhom, E Solomon, and JJ DiCarlo. 2010. Human versus machine: comparing visual object recognition systems on a level playing field. *Cosyne Abstracts* (2010).
- [37] Mary C Potter. 1976. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory* 2, 5 (1976), 509.
- [38] Pau Rodriguez, Guillem Cucurull, Josep M Gonfau, F Xavier Roca, and Jordi González. 2017. Age and gender recognition in the wild with deep attention. *Pattern Recognition* 72 (2017), 563–571.
- [39] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*. 10–15.
- [40] Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. 2017. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.
- [41] Caifeng Shan. 2010. Learning local features for age estimation on real-life faces. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*. 23–28.
- [42] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).
- [43] Simon Thorpe, Denis Fize, and Catherine Marlot. 1996. Speed of processing in the human visual system. *nature* 381, 6582 (1996), 520–522.
- [44] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [45] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 24–25.
- [46] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. 2018. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 606–624.
- [47] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*. Springer, 506–523.
- [48] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2021. A Study of Face Obfuscation in ImageNet. *arXiv preprint arXiv:2103.06191* (2021).
- [49] Juha Ylioinas, Abdenour Hadid, and Matti Pietikainen. 2012. Age classification in unconstrained conditions using LBP variants. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 1257–1260.
- [50] Minghui Zhang, Yumeng Liang, and Huadong Ma. 2019. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 151–156.
- [51] Peng Zhang, Tony Thomas, Sabu Emmanuel, and Mohan S Kankanhalli. 2010. Privacy preserving video surveillance using pedestrian tracking mechanism. In *Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence*. 31–36.
- [52] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5810–5818.
- [53] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.