2024 33rd IEEE International Conference on
Robot and Human Interactive Communication (RO-MAN)
August 26-30, 2024. Pasadena, California, USA.

2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN) | 979-8-3503-7502-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/RO-MAN60168.2024.10731324

Performing Human Shadow Detection for Camera-Based Privacy-Preserving Human-Robot Interactions

Yuhan Hu*, Prishita Ray* and Guy Hoffman

***Abstract*— Home robots are envisioned to provide in-home assistance for older adults and other people who may need help with daily tasks. To gather information for inferring user status, robots typically require cameras to detect human subjects, track their positions, and recognize their activities or poses. However, having cameras in personal spaces, such as homes, could pose privacy concerns and risks due to the potential misuse or compromise of personal image data. It can also lead to psychological unease and feelings of insecurity, stemming from the fear of being watched and recorded. To address this issue, this paper proposes a method for preserving privacy based on physically obstructing the robot's camera image and computer vision methods for detection and tracking of humans in these obstructed images. We present a hardware platform that includes a semi-transparent physical layer in front of the robot's cameras to obtain privacy-preserving shadow images, and a software framework that uses a pre-trained EfficientNet, retrained with a newly-collected dataset of human shadow images for detecting and tracking human subjects. The testing results reveal that the network achieves reliable accuracy in detecting humans from various distances and angles, and it can be applied to a new subject that it has never seen before. Finally, the algorithm is implemented in a gaze-based human-robot interaction scenario, demonstrating its ability to track humans in real time while preserving privacy.**

I. Introduction

This paper proposes and evaluates a method for preserving privacy in human-robot interactions by using physical filters in front of a robot's camera, and detecting and tracking humans in the resulting low-fidelity shadow images.

The prevalent use of cameras by robots to detect users and recognize a user's state or intention is crucial for facilitating effective human-robot interactions. Yet, when these robots are integrated into privacy-sensitive settings like homes, the presence of a camera may compromise the user's privacy, consequently diminishing the comfort level during interactions. Current technologies primarily focus on using software-related methods to protect visual privacy. These methods include modifying the captured visual content through techniques like blurring faces, altering identifiable features, or transforming the data into non-detailed forms [1], [2], [3]. However, these measures still involve the initial collection of high-resolution images, leaving a door open to potential privacy breaches through hacking or software errors. Moreover, this strategy often leaves users with minimal control over their privacy and possibly leaving them uninformed about any invasions of their privacy.

To address this challenge, we propose a privacy-centric alternative for camera-based interactions through physical filtering. By equipping the robot's "eyes" with a translucent layer, it can still access essential interaction data, via user shadows, rather than detailed images, balancing information gathering and privacy preservation, as illustrated in Fig. 1. This approach involves physically integrating privacy-protective layers (manifested as the robot's glasses), granting users complete control and awareness over their privacy settings by physically putting on the glasses. Unlike existing software-based methods, our proposed method captures the low-fidelity images from the start, avoiding the risks of privacy breaches due to software malfunctions or data leaks. Throughout the paper, we refer to the low-resolution, blurry image captured via physical filtering as a "shadow image". The detection we implemented enables the robot to detect users and track their positions based on their shadows, potentially using this information for meaningful interaction activities. One example is identifying crucial interaction events, such as a user's accidental falls, in applications like elderly care assistance.

Fig. 1: The robot wearing a pair of privacy-preserving glasses can filter high-resolution images into user shadow-captures, thus protecting users' visual privacy during privacy-sensitive tasks, such as changing clothes.

The privacy preservation approach consists of two main components: a hardware mechanism for shadow filtering, and an algorithm designed to analyze shadow images, enabling the detection and tracking of human subjects and their movements. To develop the hardware for collecting shadow images, we used the anthropomorphic Reachy robot [4], equipped with two co-mounted cameras and fitted with a pair of glasses made from translucent materials to obtain shadow-form blurry images. Our method for detecting human subjects and track their positions through shadow images is based on our previous work ShadowSense [5], where a similar approach was used to classify touch gestures in inflatable and translucent robots. In this work, however, we use a different neural network architecture, EfficientNet [6],

*These authors contributed equally to this work
All authors are affiliated with Cornell University, Ithaca, New York, USA

and apply transfer learning to retrain human detection using self-collected shadow images as training data on a pre-trained EfficientNet-v2.

To train our model, we collected a training dataset of 1182 pairs of clear and shadow images from four human subjects and from ten different head orientations of the robot. This dataset was automatically labeled using a YOLO model [7], which predicted the bounding boxes on the clear images for generating corresponding annotations for the associated shadow images. The model was trained for 8 epochs with PyTorch on a cloud-based GPU.

To evaluate the model's performance, we constructed three test datasets: the first includes a familiar human subject that the algorithm has seen before, dressed in attire that blends with the background (108 images); the second comprises a new subject the model has not previously encountered (110 images); and the third involves two known subjects moving across different distances and angles to analyze the effect on detection accuracy. The model demonstrated a detection rate of 82.4% (IoU score = 0.663) on the first dataset and 83.6% (IoU score = 0.562) on the second, indicating its reliable capability to recognize both unseen subjects and subjects wearing different outfits.

The third test set was comprised of 800 images, featuring 40 distinct distance-angle combinations from two subjects. The model achieved an average of 94.45% detection accuracy and 0.6845 IoU score, indicating a reasonable detection accuracy for all head orientation angles. The results also revealed a slight decrease in accuracy of IoU score as the distance between the robot and the human increases.

Using the model, we demonstrate the scenario of preserving users' privacy in a gaze-based human-robot interaction. Upon wearing translucent glasses, a robot engages in a privacy-maintaining mode while being able to engage in an interaction scenario by tracking the user's position, gazing at the user and using its antenna to respond based on the user's distance.

We thus show that, through the filters, robots can still process shadow images to detect a user's presence and track their movements. The method can be applied to various robot embodiment and interaction scenarios, such as performing gaze-like behaviors or detecting accidents.

II. Related Work

In this section, we discuss the prevailing challenges related to privacy concerns in camera-equipped home robots and review existing methods for visual privacy protection. Following that, we examine current models for human tracking in computer vision and machine learning that are potentially applicable to human shadow images, providing a foundation for our work.

A. Privacy Protection in Camera-Based Interactions

Robots operating in the real world, especially in privacy-sensitive areas like homes, can raise significant privacy concerns due to their potential to gather data [8], [9]. A large number of these robots employ cameras to understand human states and intentions, which is essential for effective human-robot interaction (HRI). Yet, this very feature can be a primary source of potential privacy breaches.

Most current research has tried to mitigate these concerns by post-processing visual data. For instance, existing methods use image encryption to conceal the privacy-sensitive region of interest, such as performing face de-identification to alter the face of a person in such a way that it cannot be recognized [10]. Image filtering can use also common filters such as blurring (e.g., applying a Gaussian function) and pixelating [1], [2], [3]. Some models can perform visual abstraction and object replacement to substitute the person or object appearing in the image by a visual abstraction that protects the privacy of an individual while enabling activity awareness [11]. These solutions are not fail-proof. They remain vulnerable to software failures or hacking attempts. Furthermore, software-based protection does not guarantee users full control over their privacy settings, nor does it directly notify them about the status of their privacy protection. For instance, users might remain unaware of privacy breaches resulting from hacking or software failures. Similar to our approach, other research has investigated recognizing activities in extremely low-resolution images for privacy protection, such as in [12] and [13]. However, they still capture the raw images of the user and use software-based post-processing to blur those images, which is still prone to software failures and lack of user control.

Other privacy protection alternatives use low resolution sensors or non-visual sensors, such as infrared, ultrasonic, or radar to reduce privacy risks while still enabling the detection of the environments [14], [15], [16]. For example, Tateno et al. [14] present a fall detection method using a low-resolution infrared array sensor by applying convolutional neural networks. However, these methods necessitate integrating a new sensor to replace the camera, are specific to low-resolution detection, and cannot shift between privacy preferences without switching sensors. In contrast, the proposed method allows for an easy switch of privacy modes by simply physically displacing the filters.

Drawing inspiration from our prior research [5] that enabled robots to capture users' contact shadows (notably of the hand and arm) during touch interactions and to detect touch gestures, this paper uses a similar approach to detect users' full-body shadows from a distance. In fact, previous research [5] suggested physically covering the robot's camera with a translucent material to allow the capture of users' shadows as a data source, but it primarily provided this as a conceptual framework to be explored in future work. Building on this preliminary concept, this paper aims to develop a comprehensive algorithmic design to infer user information from the full-body shadows. We also use a different neural network architecture than the one previously proposed for touch recognition.

B. Algorithms for Human Shadow Detection

To extract user information, such as identifying presence and spatial locations, it's essential to develop an algorithm

capable of analyzing shadow images to identify and locate human shadows. Below we review the state-of-art developments in deep learning and computer vision technologies that form the foundation of our methodology. The use of Deep Learning, specifically using CNN (Convolutional Neural Networks) architectures has been used to generate bounding boxes for object detection applications such as in video surveillance and robotics [17]. Optimization of the predicted bounding boxes becomes crucial for real-time object detection such as human detection, especially under noisy and occluded scenes, as discussed in [18]. YOLOv3 (You Only Look Once) [7], a one-stage algorithm from the region proposal network family, coupled with data augmentation employs logistic regression to predict bounding box coordinates and incorporates feature pyramid networks in its DarkNet [19] architecture. These enhance its robustness and accuracy in predicting bounding boxes for detection applications beyond that of traditional CNN architectures. YOLOv3 is utilized to provide reliable ground truth bounding box annotations for objects detected in image frames.

A more recent architecture, EfficientNet [6], can give near state-of-art performance with a much smaller network size and is computationally cheaper, compared to many two-stage detectors. [20] presents a survey of many modern deep learning-based object detection models, such as VGGNet (Visual Geometry Group Network) [21], CSPNet (Cross Stage Partial Networks) [22] and also describes the benefit of EfficientNet over the other architectures. Therefore, the EfficientNet model proves to be a suitable choice to serve as an object detection model on resource-constrained devices such as robots.

Transfer learning has recently gained momentum to transfer learned network weights from commonly used datasets such as COCO (Common Objects in Context) [23] or ImageNet [24] to custom datasets, which aids in faster learning and weight adjustments in comparison to training the networks from scratch. For example in [25], the *HeadPoseImageDatabase* and *HeadPoseAnnotationDatabase* [26] were used to pre-train a CNN architecture, which was then re-trained on a custom database though transfer learning for behavior recognition in real-time videos. This paper bases on the EfficientNet model and transfer learning techniques, using ground truth annotations generated from a robust model like YOLO, to train a model suitable for robot deployment such as EfficientNet on a custom-collected dataset for real-time human shadow detection. Recent advancements in transformer technology [27], [28] have demonstrated efficient and effective human detection and activity recognition, even with zero-shot learning. Future research will explore additional models to further enhance the shadow detection capabilities.

III. Method

In this section, we present the hardware and software implementation of the privacy-preserving method. This encompasses the hardware integration of semi-transparent glasses and the software framework for training the model for human detection and tracking.

Fig. 2: The Reachy robot used in the interaction wears a pair of glasses with customized privacy-preserving lens.

A. Privacy-Preserving Glasses Setup

To incorporate a shadow-filtering mechanism into the robot cameras, we designed lens consisting of two layers of transparent acrylic with a semi-transparent silicone layer sandwiched between them. This silicone layer was fabricated using a conventional mold-casting technique with a mold created from 3D printing. Empirical tests revealed that the thickness of the silicone layer directly influences the shadow's clarity: a thinner layer results in clearer images post-filtration. After preliminary tests, we opted for a silicone layer (Ecoflex 00-30) of 1mm thickness to balance between data clarity and privacy preservation. These lenses were incorporated into a pair of 3D-printed glasses, as shown in figure 2, that can be customized to fit various robot designs.

B. Software Architecture for Human Shadow Detection

The software architecture of the shadow image processing is based on neural networks that are retrained on a self-collected shadow image dataset. These networks take the shadow images as input and output the predicted bounding boxes of the human subjects. To reduce the annotation effort, we employ an automatic annotation method to generate ground truth bounding boxes for the training dataset. This involves using the YOLO model to annotate the paired clear images and then transforming these annotations for use with the shadow image dataset. In the following sections, we detail the data collection procedure for the training dataset, the automatic annotation method, and the application of transfer learning to the EfficientNet model.

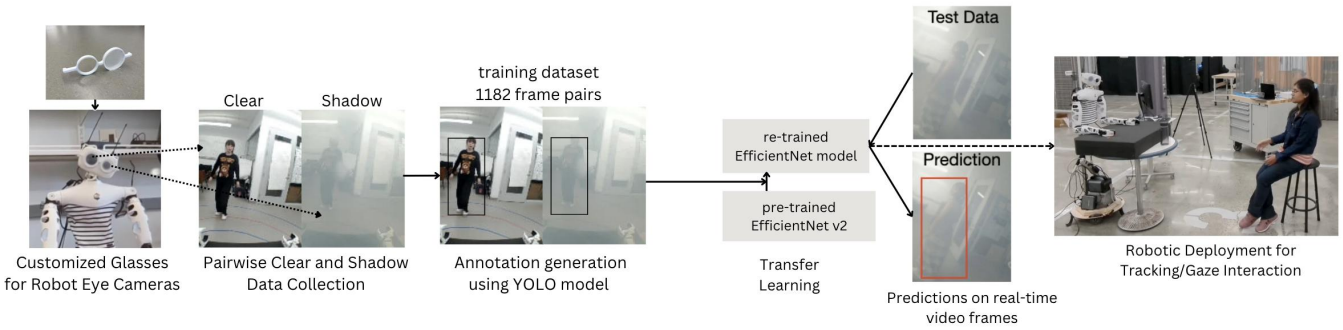

Fig. 3: Software pipeline for Human Shadow Detection, including data collection, automated annotation, transfer learning, testing, and robotic deployment.

1) Data Collection and Annotation: We use the Reachy robot's co-mounted dual-camera system. Each camera is a 1080p@30fps device with motorized zoom (FOV 65° to 125°), and the cameras are 2 inches apart, similar to the natural human eyes separation.

For data collection, we covered the left camera with fully transparent acrylic glass, while the right camera was covered with a semi-transparent shadow filter, as described in Section III-A. The training data were collected through this bi-camera system, with each shadow image paired up with the corresponding clear image, as illustrated in figure 3. A coordinate transformation through linear shifting was performed to align the two cameras' coordinate systems to compensate for the 2-inch linear separation, so that the clear and shadow images were captured from the same direction through the eye cameras.

The training dataset consists of 1182 paired shadow-clear image frames. We collected images to cover a variety of human profiles with posture variances, distances from the robot, different robotic configurations, and environmental conditions. We included a total of four human subjects (two male and two female) as volunteers to participate in the data collection. During the data collection activities, each subject was asked to perform several static and dynamic daily activities in front of the robot, such as pulling out a chair and sitting, drinking from a mug, reading a book, walking around, working on a laptop, dancing, reaching for items, etc. They are also asked to move out of the robot's view from time to time to include scenes without human presence. The robot was posed with ten different head orientations facing different angles in space. Participants were positioned within a distance range of four to twelve feet from the robot. The data collection was performed over several sessions, at different times of the day with varying lighting conditions and slight variations in the background.

We collected paired video data recorded across four participants, each recorded for around five minutes. The paired shadow-clear video data were pre-processed to construct individual image frames at a rate of one frame per second to construct shadow/clear image datasets. To improve training efficiency, the image size was reduced from the original size of $480 * 640$ from the robot camera to $120 * 160$.

To automatically annotate the shadow image frames, we used the YOLO model [7] to generate ground truth bounding boxes on the paired clear image frames. The YOLO model, which was pretrained on the large-scale object detection and segmentation COCO dataset, includes the human subject as one of the object categories. One advantage of using the COCO dataset is that it has more instances per category compared to other datasets, such as ImageNet-1k [24], which aids in learning more detailed object models capable of precise 2D localization [23].

We input the clear images into the YOLO model to output the bounding boxes of the human subject. This output was processed with the camera coordinate transformation and used as the ground truth bounding box annotation for the paired shadow images to train the EfficientNet, as described in the following section. We also validated the accuracy of the automatic annotation method by comparing the annotation labels to manually labeled sample data, and the results are reported in Section IV-A.

2) Transfer Learning with EfficientNet: To detect and track human shadows, we employ object-based deep learning neural networks that draw bounding boxes around human shadows, thereby inferring the presence and physical position of humans in the environment. We used the pre-trained EfficientNet-v2 as our foundation model and applied transfer learning to retrain it with a self-collected shadow image dataset, enhancing its performance for the custom use case.

The choice of EfficientNet-v2 over other state-of-the-art models is based on the following reasons: (1) EfficientNet-v2 is a Convolutional Neural Network architecture that boasts high performance for bounding box prediction and is memory-efficient, making it suitable for storage and use on resource-constrained devices like the Reachy robot, in addition to offering fast prediction speeds. (2) It also features rapid training speeds and relatively lower number of parameters, being almost 6.8 times smaller than state-of-the-art models such as YOLO [19] and RetinaNet [29].

The structure of EfficientNet-V2, from input to output, consists of a Convolution layer, followed by three Fused-MBConv (Mobile Inverted Bottleneck Convolution) blocks, three MBConv blocks, and a final block combining convolution, pooling, and fully connected layers [30]. Table I pro-

vides details of the network architecture. The model employs progressive learning through adaptive regularization, incrementally increasing both the image size and regularization during training.

We utilized the open-source PyTorch-Lightning [31] machine learning framework to implement the model, training it on a GPU-equipped server with 14 VCPU and 1 GPU. Through transfer learning, we adapted the original EfficientNet-v2 model to fit our self-collected shadow image dataset with automatically-generated bounding boxes for annotations. The model was trained using the Adam optimizer with a learning rate of 0.0002 and mean Intersection over Union (mIoU) as the minimization objective for the target and predicted bounding boxes, setting a threshold of 0.44 and class confidence prediction of 0.2. At each epoch, we shuffled the shadow dataset, using 50% of the frames for training with their corresponding ground truth annotations and 50% for validation. A small batch size of 2 was chosen, as it is preferable for an object-detection network dealing with images of lower resolution ($120*160$ pixels) and because the EfficientNet-v2 network performs computationally intensive operations, particularly due to the MBConv and Fused-MBConv layers. Loss convergence was observed after just 2-3 epochs. To prevent overfitting, the model was trained for a total of 8 epochs. Figure 3 depicts the transfer learning pipeline for training the EfficientNet-v2 model.

Stage	Operator	Stride	# Channels	# Layers
0	Conv3X3	2	24	1
1	Fused-MBConv1, k3X3	1	24	2
2	Fused-MBConv4, k3X3	2	48	4
3	Fused-MBConv4, k3X3	2	64	4
4	MBConv4, k3X3, SE0.25	2	128	6
5	MBConv6, k3X3, SE0.25	1	160	9
6	MBConv6, k3X3, SE0.25	2	272	15
7	Conv1X1+Pooling+FC	-	1792	1

TABLE I: EfficientNet-V2 architecture [30]

IV. Model Evaluation

We began with a validation of the automated annotation generation method. Then, we evaluated the performance of the retrained EfficientNet on three test datasets, which were collected using unseen human subjects at various distances and with different configurations of the robot's head, to assess the model's generalization capabilities.

A. Validation Results of Automated Annotation Generation

To assess the reliability of the ground truth bounding box annotations automatically generated by the YOLO model, we performed a validation test for the automatically generated bounding boxes. We manually annotated a small sample data set and compared the manually annotated bounding boxes to the automatically generated bounding boxes.

We firstly performed a detection rate validation to identify human subjects present in image frames. We randomly sampled 200 image frames from the training dataset, repeating this process five times. Some frames contained human subjects, while others did not. Across all five samplings, totaling 1,000 image frames, the YOLO model successfully detected all human subjects when present.

We then compared the automatically annotated bounding boxes with the manually labeled ones by constructing a validation test set of 79 image frames. These frames, randomly selected from those containing human subjects, showed that the automatically annotated bounding boxes achieved a mean Intersection over Union (IoU) score of 0.80 when compared to the manually labeled ones.

Although this is considered a satisfactory score, there is still a slight deviation from the manually annotated ground truth in terms of IoU score, and that may have introduced noise into the training set. However, such noise can potentially aid in the generalization of the model's performance during training. Neural networks can find learning from small datasets challenging, as they may end up memorizing the examples. Introducing noise during training can enhance the robustness of the training process and decrease the generalization error.

B. Evaluation of Human Shadow Detection Accuracy

To evaluate the accuracy and robustness of the retrained EfficientNet-v2 in performing human shadow detection and tracking, we constructed three test datasets. These datasets were designed to assess detection accuracy under various conditions. Test Set 1 includes a human subject previously seen in the training dataset but wearing a different outfit with a color similar to the background. Test Set 2 features a subject not seen during training. Test Set 3 involves two subjects from the training dataset wearing different outfits and introduces variations in the subjects' distances from the robot and angles of the robot's head orientation, aiming to test detection accuracy in relation to spatial dynamics.

We employed several metrics to evaluate detection accuracy: **detection rate**, to measure the proportion of correctly predicted image frames; **precision**, to identify the model's false positive predictions; **recall**, to capture the model's false negative predictions; and **mIoU** (mean Intersection over Union), which quantifies the overlap of predicted bounding boxes relative to ground truth bounding boxes.

1) Test Set 1: Test Set 1 is comprised of a human subject who previously appeared in the training dataset but is wearing a different outfit (white) that blends with the background (white wall). This setup is designed to assess how the algorithm performs in challenging situations where shadows may blend with the environment and might not be properly captured. The set includes 108 image frames collected from the subject moving in front of the robot, featuring varied distances, robot head orientations, and slight variations in the lighting conditions of the environment. The retrained EfficientNet-v2 model achieves a detection rate of 0.824, with a precision of 1 and a recall of 0.763. The IoU (Intersection over Union) scores of the successfully detected samples are averaged at 0.663 ± 0.102 (mean $\pm$ standard deviation).

2) Test Set 2: Test Set 2 features a new human subject (male) previously unseen in the training dataset. This set aims

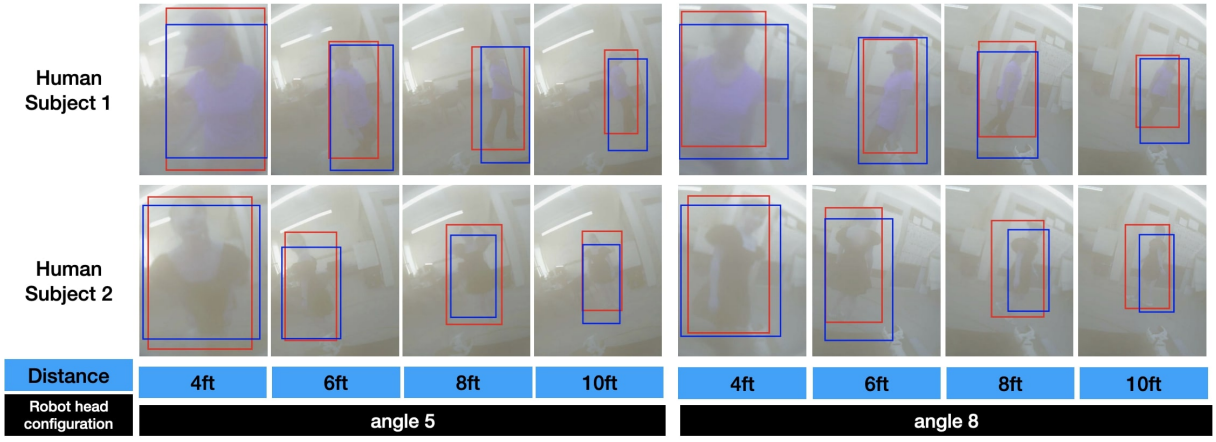

Fig. 4: Test results of the re-trained EfficientNet-v2: The ground truth (red) and predicted (blue) bounding boxes of user's positions from the shadow images, sampled from Test Set 3 with distance and robot head configuration labels.

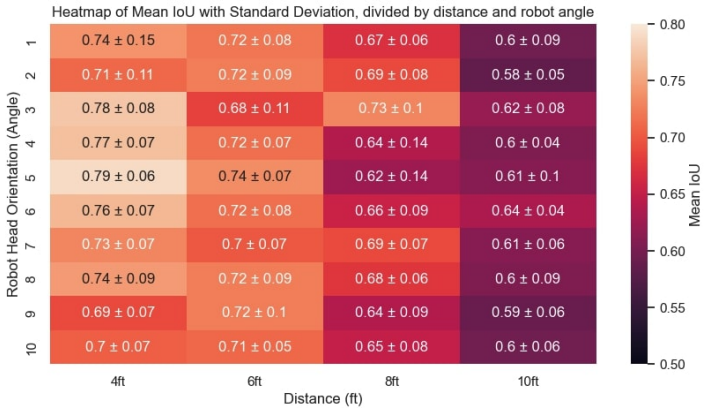

Fig. 5: Average mIoU scores with distance (ft) and angles for two test subjects from Test Set 3.

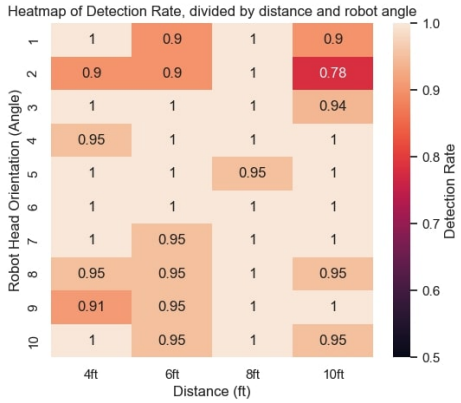

Fig. 6: Average detection rates with distance (ft) and angles for two test subjects from Test Set 3.

to assess the model's ability to generalize to a new human subject in a real-world scenario. We captured 110 image frames of this new test subject, capturing their movements in front of the robot at various distances and with differing robot head orientations. The model achieved a detection rate of 0.836, with a precision of 0.890 and a recall of 0.875. The Intersection over Union (IoU) scores for the successfully detected samples averaged at 0.5624 ± 0.218.

3) Test Set 3: Test Set 3 includes two subjects previously seen in the training dataset, but with different outfits. The dataset features images that are collected and categorized based on four varying distances between the human subjects and the robot (4 ft, 6 ft, 8 ft, 10 ft), as well as ten distinct angles of the robot's head orientation (ranging from -0.5 to +0.5 radians, with intervals of 0.1 radians), starting from the robot's standard neutral position (0 radian meaning facing directly forward). In total, 800 image frames were collected, with each subject being captured in 10 frames for every distance and head orientation combination. This aims to evaluate the model's performance by considering spatial relationships, including distance and head angle, as key factors in capturing the physical relationship between humans and the robot.

Overall, the model has achieved an average detection rate of 0.9445, with precision of 1 and recall of 0.967. The average Intersection over Union (IoU) scores are 0.6845 ± 0.0834. Figure 6 captures the average detection rate for each distance and head orientation combination; where as figure 5 depicts the mean IoU score and standard deviation for each distance and head orientation combination. Figure 7 further captures the accuracy-distance relationship, where as the distance between humans and the robot increase, the mean IoU scores decrease for both test human subjects. In sum, as the algorithm can achieve relatively high detection accuracy, the accuracy (mIoU) is affected by the distance between the human and the robot, whereas the robot's head orientation does not affect the accuracy in a noticeable way.

V. Privacy-Protected Human-Robot Gaze Interaction

To illustrate the practical application of our method in real-world human-robot interaction scenarios, we integrate a privacy-preserving interaction scenario wherein the robot

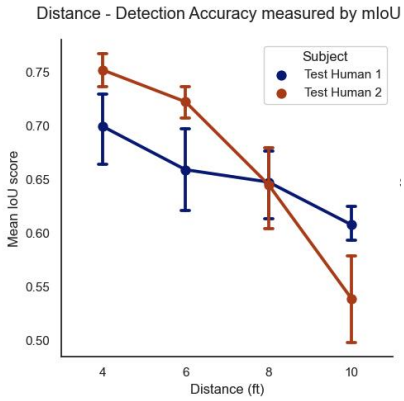

Fig. 7: Average mIoU scores with distance (ft) for two test subjects from Test Set 3

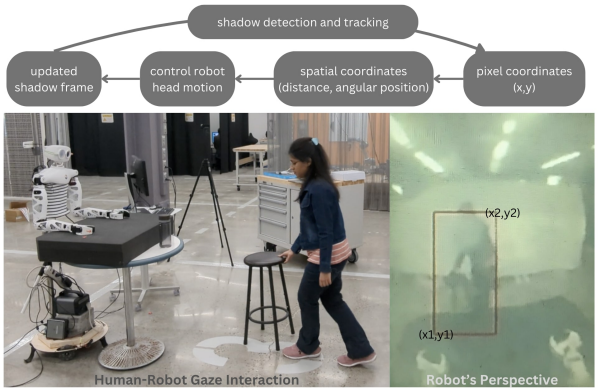

Fig. 8: Integration of shadow tracking in human-robot gaze interaction.

tracks the user and "gaze" at the user in real-time. Below, we detail the interaction design of the robot's behavior and the implementation within the robot control system.

A. Interaction Design of Robot Behaviors

To illustrate the interactions facilitated by the privacy-preserving algorithms, we have designed an interaction scenario that necessitates visual information gathering while emphasizing the importance of privacy preservation. We imagined the interaction of human tracking behavior within a home setting for social scenarios, where the robotic "gaze" employs human-like nonverbal communication. This human tracking behavior can demonstrate the robot's attention, awareness of the user, signal interest and engagement, and make the interactions feel more natural and intuitive. However, human tracking behavior could potentially lead to discomfort if users feel they are being stared at constantly. This discomfort may stem from low trust in the robot and concerns over the leakage of their personal visual data captured through gaze interaction. Thus, we aim to utilize the system to enable human tracking and gaze interaction while preserving user privacy and increasing users' psychological comfort during the interaction.

We utilize the Reachy robot platform [4] and its built-in actuators and controller for behavior generation. This includes its neck with 3 degrees of freedom (DOF) and two antennas, each with one DOF. We use neck movement to orient Reachy's 'eyes' to always face the human, and the antennas to reflect the robot's internal state in relation to the physical distance between the human and the robot, as this distance may indicate the level of interaction engagement and interest. Specifically, the behaviors are defined as follows: Upon detecting a human, Reachy rotates its head to gaze towards the user, simultaneously shifting its antennas slowly to express human recognition. When the robot detects that a human subject's distance is within its close proximity, it shakes its antennas to express "excitement". In the absence of a human, the robot maintains a static gaze. A supplementary video demonstrates the interaction of the human tracking behavior.

B. Integration with Robot Control System

Below, we describe the control algorithms designed to execute human tracking behavior upon receiving the human tracking bounding box from the detection algorithm, enabling real-time gazing at the user.

The robot is equipped with high-resolution cameras capable of capturing images up to 1080p at 30 fps. We activate the built-in auto-focus function when wearing the privacy-preserving glasses to control the clarity of the incoming video feed, helping the robot clearly capture the user's shadow. Upon capturing the real-time frame of the shadow image, it is fed into the re-trained EfficientNet model to predict the bounding box coordinates of the detected human subject in the image. We annotate $(x1, y1)$, and $(x2, y2)$ as the diagonal pixel coordinates of the predicted bounding boxes. The spatial position of the human (D, θ) in relation to the robot - where D denotes distance, and the θ denotes angular position from the robot's gazing direction - is then derived based on the image coordinates through a pre-calibrated camera-space transformation matrix.

We employed a proportional control strategy to quantify the discrepancy between the robot's built-in *'look at'* function and the observed human spatial coordinates in relation to the robot. We fine-tuned the control parameter K iteratively to achieve optimal tracking to real-time detection. We use feedback from the camera's image data to update the human's current position relative to the robot with proportional control until the person is positioned in the center of the robot's perspective, within a certain threshold. As a result, the control algorithm achieves real-time human tracking at a rate of approximately 1 fps, taking into account the delay from the visual prediction and the head movement.

In addition, the distance between the human and the robot D is derived from the proportional size of the bounding box $x2 - x1$ and $y2 - y1$ in relation to the size of the image frame (W and H). When the robot detects that the subject's distance is within $2ft$, the antenna motors are actuated to oscillate back and forth three times to express 'excitement.' To avoid control conflicts, both the tracking function and head motion are temporarily paused during the

antenna motion. Figure 8 depicts the control pipeline in the interaction scenario.

Conclusion

In this paper, we introduced a vision-based, privacy-preserving method for human-robot interaction, designed specifically for robots interacting within personal spaces. We describe our hardware setup and software architecture for tracking human shadows using a retrained EfficientNet-v2 model, utilizing self-collected and automatically annotated shadow images. The model is evaluated across three test sets, demonstrating reasonably high detection accuracy for unseen human subjects, and subjects that may blend into the background color. Additionally, we assess the detection accuracy across various distances and robot head orientations, noting a slight decline in accuracy as the distance between the user and the robot increases. Lastly, we demonstrate the application of our method in human tracking behavior, enabling real-time tracking and gazing with the user. For future work, we aim to conduct user studies to better understand the interaction experience and user perceptions of the privacy-preserving detection method in real-world scenarios.

ACKNOWLEDGMENT

This research was supported under NSF National Robotic Initiative Award NRI:1830471.

References

[1] M. Rueben and W. D. Smart, "Privacy in human-robot interaction: Survey and future work," *We robot*, 2016.

[2] P. Korshunov, A. Melle, J.-L. Dugelay, and T. Ebrahimi, "Framework for objective evaluation of privacy filters," in *Applications of Digital Image Processing XXXVI*, vol. 8856. SPIE, 2013, pp. 265–276.

[3] 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)*. IEEE, 2016, pp. 1–9.

[4] "Reachy, an open source programmable humanoid robot," Pollen Robotics. [Online]. Available: https://www.pollen-robotics.com/

[5] Y. Hu, S. M. Bejarano, and G. Hoffman, "ShadowSense: Detecting human touch in a social robot using shadow image classification," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–24, 2020.

[6] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[7] H. R. Aradhya *et al.*, "Object detection and tracking using deep learning and artificial intelligence for video surveillance applications," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, 2019.

[8] M. E. Kaminski, M. Rueben, W. D. Smart, and C. M. Grimm, "Averting robot eyes," *Md. L. Rev.*, vol. 76, p. 983, 2016.

[9] K. Caine, S. Šabanovic, and M. Carter, "The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 343–350.

[10] B. Meden, M. Gonzalez-Hernandez, P. Peer, and V. Štruc, "Face deidentification with controllable privacy protection," *Image and Vision Computing*, vol. 134, p. 104678, 2023.

[11] T. Webb, S. S. Mondal, and J. D. Cohen, "Systematic visual reasoning through object-centric relational abstraction," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] M. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[13] M. U. Kim, H. Lee, H. J. Yang, and M. S. Ryoo, "Privacy-preserving robot vision with anonymized faces by extreme low resolution," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 462–467.

[14] S. Tateno, F. Meng, R. Qian, and Y. Hachiya, "Privacy-preserved fall detection method with three-dimensional convolutional neural network using low-resolution infrared array sensor," *Sensors*, vol. 20, no. 20, p. 5957, 2020.

[15] M. Ogata, S. Murakami, T. Mikura, and I. E. Yairi, "Privacy-preserving monitoring system with ultra low-resolution infrared sensor." in *AI4Function@ IJCAI*, 2020, pp. 26–32.

[16] S. Mashiyama, J. Hong, and T. Ohtsuki, "Activity recognition using low resolution infrared array sensor," in *2015 IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 495–500.

[17] A. Pathak, M. Pandey, and S. Rautaray, "Application of deep learning for object detection—— procedia comput," *Sci*, 2018.

[18] E. U. Haq, H. Jianjun, K. Li, and H. U. Haq, "Human detection and tracking with deep convolutional neural networks under the constrained of noise and occluded scenes," *Multimedia Tools and Applications*, vol. 79, pp. 30 685–30 708, 2020.

[19] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia computer science*, vol. 199, pp. 1066–1073, 2022.

[20] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, p. 103514, 2022.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[23] 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[24] 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.

[25] T. Kujani and V. D. Kumar, "Head movements for behavior recognition from real time video based on deep learning convnet transfer learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 7047–7061, 2023.

[26] U. Weidenbacher, G. Layher, P.-M. Strauss, and H. Neumann, "A comprehensive head pose and gaze database," in *2007 3rd IET International Conference on Intelligent Environments*. IET, 2007, pp. 455–458.

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[28] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[29] M. Cheng, J. Bai, L. Li, Q. Chen, X. Zhou, H. Zhang, and P. Zhang, "Tiny-RetinaNet: a one-stage detector for real-time object detection," in *Eleventh international conference on graphics and image processing (ICGIP 2019)*, vol. 11373. SPIE, 2020, pp. 195–202.

[30] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," in *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 463–469.

[31] K. Sawarkar, "Deep learning with PyTorch Lightning: Swiftly build high-performance artificial intelligence (ai) models using python." Packt Publishing Ltd, 2022.