Impacts of Image Obfuscation on Fine-grained Activity Recognition in Egocentric Video

Soroush Shahi[1,3], Rawan Alharbi[1,3], Yang Gao[1,3], Sougata Sen[4],
Aggelos K Katsaggelos[1,2], Josiah Hester[1,2,3], Nabil Alshurafa[1,2,3]
[1]Department of Computer Science, [2]Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA
[3]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA
[4]Department of Computer Science and Information System, BITS, Pilani, Goa, India
{soroush, rawan.alharbi, yang.gao, sougata.sen, a-katsaggelos, josiah, nabil}@northwestern.edu

2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 978-1-6654-1647-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/PerComWorkshops53856.2022.9767447

***Abstract*—Automated detection and validation of fine-grained human activities from egocentric vision has gained increased attention in recent years due to the rich information afforded by RGB images. However, it is not easy to discern how much rich information is necessary to detect the activity of interest reliably. Localization of hands and objects in the image has proven helpful to distinguishing between hand-related fine-grained activities. This paper describes the design of a hand-object-based mask obfuscation method (HOBM) and assesses its effect on automated recognition of fine-grained human activities. HOBM masks all pixels other than the hand and object in-hand, improving the protection of personal user information (PUI). We test a deep learning model trained with and without obfuscation using a public egocentric activity dataset with 86 class labels and achieve almost similar classification accuracies (2% decrease with obfuscation). Our findings show that it is possible to protect PUI at smaller image utility costs (loss of accuracy).**

***Index Terms*—Human Activity Recognition, Wearable Camera, Deep Learning, Image Obfuscation**

I. Introduction

Hand-related human activity recognition (HAR-2) plays a major role in several applications including remote assistance and human-robot interaction, and is critical to understanding factors that influence health outcomes [1] such as dietary monitoring [2]. Leveraging self-report and natural observation for HAR-2 is burdensome and costly, negatively impacting scalability and feasibility when deployed in longitudinal studies [2]. Needed are automated methods for HAR-2. To this end, wearable cameras are used to automate the monitor of HAR-2 and provide both precise timing and visual confirmation of these activities [3]. Recent progress in computer vision algorithms have considerably enabled the use of wearable cameras for automated HAR-2 in real-world settings.

Despite the popularity of wearable cameras in research and advancements in automated detection methods, RGB video data inherently captures data that people are uncomfortable sharing, reducing people's willingness to wear cameras [4]. This poses a major challenge for health researchers intending to capture naturally occurring behavior, and computer scientists intending to publish their datasets for transparency, reproducibility, and the advancement of science. Researchers often have to go through a rigorous process to de-identify data and remove sensitive information before they allow public access to datasets [5]. This usually involves either the strict control over data collection or a blocklist approach to obfuscating sensitive information (e.g., where a predefined set of labels representing sensitive objects like faces are removed). Controlling data through strictly controlled protocols prevents the ability to capture and understand naturally occurring activities in real-world settings, weakening the impact and generalizability of the study. Likewise, limiting the definition of privacy to a set of sensitive objects is not ideal, because what an individual perceives to be sensitive information is application-based and varies from one individual to the next. Finally, removing predefined labels either requires tedious manual human inspection or automatic detection methods, which adds another source of error.

To address these concerns in human activity recognition, researchers recently investigated the tradeoff between preventing the capture of personal user information (PUI) and image utility [4], [6]. In most cases, cameras capture more information than intended and more details than needed, which can result in violation of user privacy [6]. By highlighting this fact, we looked into methods applied in egocentric vision that eliminated non-essential information from the scene. Alharbi et al. show how image obfuscation methods in wearable cameras can protect PUI, but often come at the cost of utility in HAR-2. One obfuscation method shown to be most effective in protecting PUI is that of masking background pixels. This is the most extreme background obfuscation method since it completely eradicates background pixels by setting background pixels to a single color (e.g., black). Moreover, this method was shown to be non-inferior in visually confirming the wearer's activity, when it came to detecting the wearer's hand-related activities (e.g., eating/drinking, making phone calls), especially when an object was present in-hand. However, the effect of masking on algorithms that automatically detect behavior was not studied [6]. Similarly, recent literature stresses the importance of the hands and objects in HAR-2 from egocentric cameras [7], [8]. Inspired by these works, we conducted our experiments on a large public dataset released for HAR-2 and tested the effect of hand-object-based mask (HOBM) obfuscation, that is masking everything in the scene except for the hand and *active* objects (i.e., objects in-hand) on a machine learning model's ability to automatically classify fine-grained activities.

Authorized licensed use limited to: Northwestern University. Downloaded on October 14,2022 at 19:51:33 UTC from IEEE Xplore. Restrictions apply.

We list our contributions as follows:

- We test the impact of an extreme obfuscation method, masking all objects except the hand and active object (hand-object-based mask obfuscation), on hand-related human activity recognition.
- Informed by our analysis, we further study the impact of obfuscation for each action class and provide insights on the cases where obfuscation *decreases* and *increases* utility (accuracy).

Our work demonstrates that one can maintain high privacy standards with a very minimal loss of (and some cases an improvement of) utility. By this work, we aim to reinforce the principle of least privilege data privacy, by encouraging the practice of collecting only the necessary information in wearable egocentric cameras. In doing so, we provide further support in challenging the fact that privacy can only be achieved by compromising a large amount of utility[1].

II. Related Work

Prior research focused on different obfuscation methods to improve privacy by using mainly two approaches: (1) partial obfuscation (e.g, screens, plate numbers, bystanders) [9]–[11], and (2) total obfuscation (e.g, low resolution or blur) [4], [12], [13]. Traditional partial obfuscation approaches include pixelization or blurring sensitive information in the image, such as faces to prevent identification [9], [10]. Korayam et al. enhanced life logging experiences by addressing privacy concerns that were raised by screens; they detected and obfuscated desktop and laptop computer screens [14] but did not include phones, tablets, TVs, or other electronic devices. Yan et al. explored the effect of partial obfuscation on the automated detection of activity recognition using third person view camera videos collected from the internet (i.e., YouTube) and masking the pixels corresponding to humans in the images [11]. They showed that removing the human from the image reduces the accuracy relatively by 9%. Obfuscation through masking has been shown to be useful in reducing privacy concerns, but it is not known whether it remains useful in enabling the automated detection of hand-related activities in the egocentric view. The challenges of processing data and detecting activities of interest from a third-person surveillance camera view are different than that of egocentric (first person) views because of rapid camera motion (due to its position on the body), proximity to the activity of interest, and presence of distracting objects that are not of interest [15].

Adversarial training is a common technique used to prevent information exposure among different data modalities [16]–[19]. Researchers utilize this technique to address privacy concerns by anonymizing sensitive content in videos. Ren et al. proposed a video face anonymizer that uses adverserial training to remove the user's face while maximizing the activity detection performance [20]. This learning process is modeled as a fight between a video anonymizer that tries to remove privacy sensitive information while preserving enough information from the scene for the intended detection task at-hand. Although adverserial learning approaches can achieve this trade-off in a systematic way, all these approaches assume privacy is buried inside labels like faces.

Total obfuscation is another method used to enhance protecting user information without the burden of defining privacy labels. Recently, methods for activity recognition propose building extremely low-resolution images [12], [13]. Ryoo et al. show the possibility of activity recognition from extremely low-resolution images through a concept called inverse super resolution (ISR) [13]. The idea behind ISR is that multiple low-resolution images may contain an equivalent amount of information to a single high-resolution image. Dimiccoli et al. explore a study that showed how much people were willing to wear cameras according to different levels of image degradation [4]. In this study, they tested multiple total obfuscation methods using different blur intensities and found a positive relationship between the amount of image degradation and participant willingness to be captured by the wearable camera. Although total obfuscation methods like extreme low resolution or blur address privacy concerns, they are known to significantly reduce recognition accuracy [21].

Prior studies show the importance of keeping the hands and the *active* objects (objects in-hand) to enable visual confirmation of hand-related activities [6]. In contrast to existing partial obfuscation approaches, which target the sensitive information and try to remove them, our approach focuses on localizing the activity of interest (keeping hands and object in-hand) in the image and removing everything else.

For the purpose of this work, we focused on RGB data. However, prior research focused on HAR-2 from derivatives of RGB modality including optical flow or hand trajectory [22], [23]. While these approaches are not as susceptible to the privacy concerns discussed earlier, they are not always useful for recognizing large sets of activity labels, they often require unique methods and network structures, and they fail to provide visual confirmation of activities if needed (after the model runs to provide further confirmation). By using RGB data, we keep our goal accessible to a wide range of tasks and methods and provide a proof of concept.

III. Method

Figure 1 shows the overview of our framework. In this section, we describe our obfuscation method, activity recognition training and evaluation method, and the dataset used.

A. Hand-Object-Based Mask Obfuscation (HOBM)

Motivated by recent research demonstrating that the hands and the object in-hand provide enough information to distinguish fine-grained hand-related activities [7], [8], [22], HOBM masks all pixels other than the hand and object in-hand. We first detect the hand and object in-hand. Then, we apply our obfuscation function to remove unnecessary information, correct any detection errors, and finally generate the obfuscated image.

[1]Source code available: https://github.com/HAbitsLab/HOBM

Authorized licensed use limited to: Northwestern University. Downloaded on October 14,2022 at 19:51:33 UTC from IEEE Xplore. Restrictions apply.

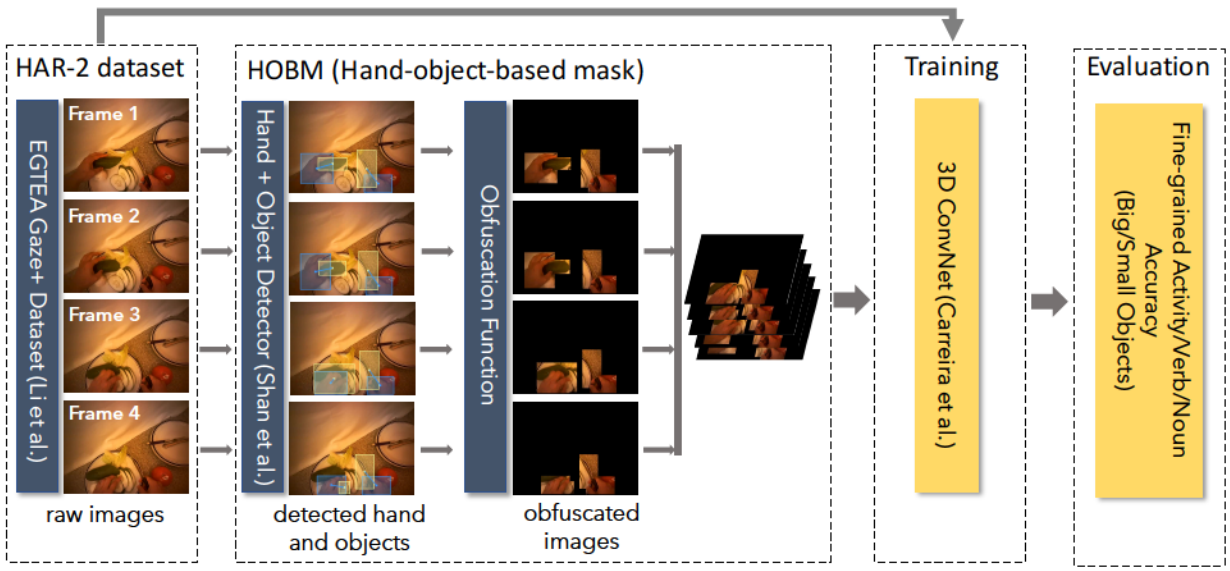


Fig. 1. Diagram of the methodological framework; Raw images from a HAR-2 dataset are passed to HOBM obfuscation block to generate obfuscated images. These images are used to train and evaluate an activity recognition model and then compare its performance with the same model trained on raw images.

1) Hand+Object Detection: While there exist several works that detect hands and objects [24]–[26], we use the model proposed by Shan et al. [27] for the following reasons: (1) it is a one-step approach where both the hands and the object in-hand are detected, and (2) cross-dataset analysis from this work demonstrates the superiority of this method with regards to generalizability and accuracy among other hand detection methods. In our obfuscation approach, we use the off-the-shelf model provided by Shan et al. as-is, without retraining or fine-tuning the model, because the model generalizes very well to other datasets [27]. We do, however, add a post-processing step to increase the hand- and object-detection precision and ensure that the obfuscation method does not reveal extra information due to detection error. We perform this by imposing a high confidence threshold (80%) for the prediction. We also restrict the area of the bounding box to minimize the detection error that exposes extra unnecessary information. Specifically, we shrink any bounding boxes with a size greater than 90% of the whole image by half (empirically selected). Given recent advances in hand detection models and edge computing, we will be able to perform this processing step on-device (e.g., TinyML), which would enable us to meet even higher privacy standards.

2) Obfuscation Function: For a single frame, let H be the set of all pixels in the image which belong to the bounding box around a participant's hand. Similarly, let O be the set of all pixels in the image which belong to the bounding box around the visible objects in the image. We define set A as a subset of O that only contains pixels belonging to the active objects, meaning that the object they represent interact with pixels related to hand pixels from set H. Finally, given the intensity function of the image for location (u, v) as $I(u, v)$, the obfuscation function is defined as:

$$\text{obf}(I(u,v)) = \begin{cases} I(u,v), & \text{if } I(u,v) \in H \cup A \\ 0, & \text{otherwise} \end{cases}$$

In the above definition, both set H and set A are generated using the hand+object detector mentioned earlier.

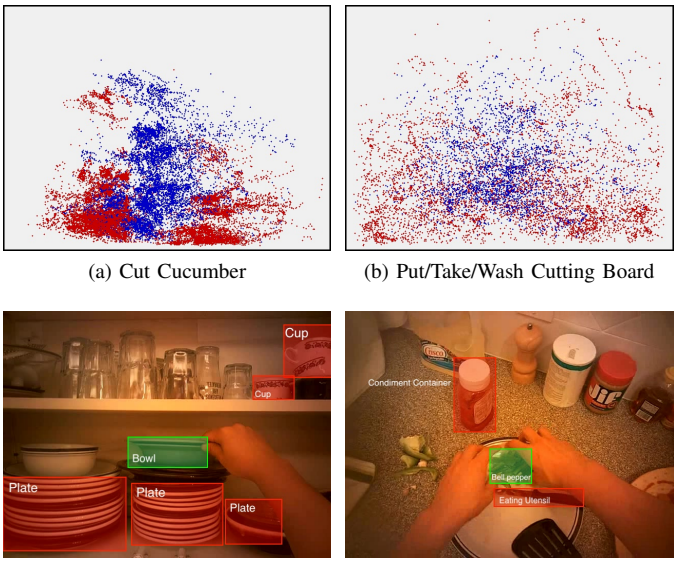
B. Dataset

We evaluated our approach on a publicly available and widely used egocentric dataset (EGTEA Gaze+ [28]) to test the impact of our obfuscation method on HAR-2 tasks. The EGTEA Gaze+ dataset was collected from a wearable camera mounted on participants' heads while the participants were preparing different recipes in a kitchen. This rich dataset contains 28 hours of fine-grained cooking activities from 86 unique sessions, collected from 32 distinct individuals. The activities combine 53 nouns and 19 verbs for a total of 106 class labels. As our obfuscation method is based on the presence of hands, and objects in the hands, we removed 16 fine-grained activity classes that deal with non-handheld objects (e.g., open fridge). In addition, we merged some overlapping classes that are even hard for a human to distinguish between (e.g., stir with cooking utensils and stir with eating utensils). This resulted in a final total of 86 class labels.

C. Activity Recognition Model Training

To test the impact of our obfuscation on the task of activity recognition, we had to train an activity detection model with HOBM obfuscation method and compare its performance with the same model trained on raw images. We used the I3D network [29] because it is the current state-of-the-art RGB-based activity recognition model for the dataset that we used. We used a pretrained ResNet50 [30] on Kinetics 400 [31] as a backbone and trained the model for 60-80

Authorized licensed use limited to: Northwestern University. Downloaded on October 14,2022 at 19:51:33 UTC from IEEE Xplore. Restrictions apply.



(a) Cut Cucumber (b) Put/Take/Wash Cutting Board

(c) Take Bowl (d) Take Bell Pepper

Fig. 2. (a)-(b): Distribution of the center of hands (red dots) and objects (blue dots) on the 2D image plane for image sequences with different object sizes. Hand appears on the scene 2.6x times more frequently when the object is small (cucumber). (c)-(d) Presence and detection of confounding objects (red bounding box) as opposed to the correct object in hand (green bounding box) can reduce model ability to detect correct activity class labels.

epochs until convergence. Learning rate schedules were used during the training to reduce learning rate at specific epochs. Finally, different temporal input sizes (e.g., 16 frames and 32 frames) were tested as a parameter to fine-tune the results. We categorize the activities by object size and report the results in Section IV.

D. Evaluation

We report both accuracy and mean class accuracy (calculates accuracy for each class and then calculates its arithmetic mean over all classes) of the model on fine-grained activity classification and verb and noun classification, both when trained and tested on raw images and obfuscated images. In addition to accuracy, we further explain the impact of object size on the results. Finally, we discuss class labels for which the impact of obfuscation is substantial.

IV. Results and Discussion

Table II shows the result of fine-graned activity classification on the EGTEA Gaze+ dataset. HOBM achieves an accuracy and mean class accuracy reduction of 2% and 4%, respectively, compared to the raw image baseline method. Table II highlights the results based on the size of the object in hand. We see no difference in accuracy using the obfuscation method when objects are small (e.g., knife or bowl), compared to a 14% reduction when the objects are large (e.g., cooking pan or pot). The confusion matrices presented in Figure 3 show how similar the results are between processing the obfuscated and raw images on small objects. Considering all object sizes, training and testing using HOBM resulted in only a 3% *relative* reduction in accuracy of HAR-2. Yan et al. [11] shows a 9% reduction in accuracy when applying a mask to human beings to maintain privacy and accuracy. Compared to prior research, our approach shows promise in its ability for HAR-2 using mask obfuscation. We also ran an experiment with the model when it is trained on raw images but tested on obfuscated ones. Results from Table I shows that a model which is exposed to more information (i.e., raw images) during training will not necessarily perform as well when tested on images with less information (i.e., the obfuscated image). This is justified by the fact that the model learns to rely on information beyond the object and activity of interest, lending greater credence to models that enhance the effect of the important parts of the image while diminishing the effect of other parts. The remainder of this section further delineates the positive and negative effects of image obfuscation on HAR-2.

TABLE I
Activity recognition accuracy of the I3D network on the EGTEA Gaze+ dataset for various object sizes, when the model is trained on raw images but tested with HOBM.

Object Size	Obfuscated?	Acc (Mean Class Acc)		
		Fine-grained Activity	Verb	Noun
Small	Yes	0.4 (0.38)	N/A	0.52 (0.53)
Large	Yes	0.57 (0.57)	N/A	0.89 (0.92)
All	Yes	0.37 (0.36)	0.57 (0.53)	0.5 (0.5)

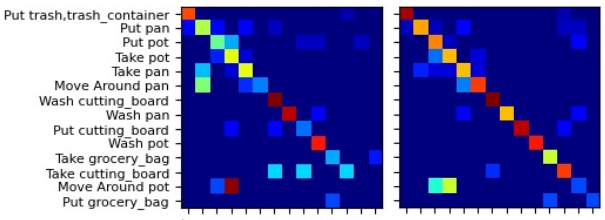
TABLE II
Activity recognition accuracy of the I3D network on the EGTEA Gaze+ dataset for various object sizes, and for HOBM applied on images containing these objects.

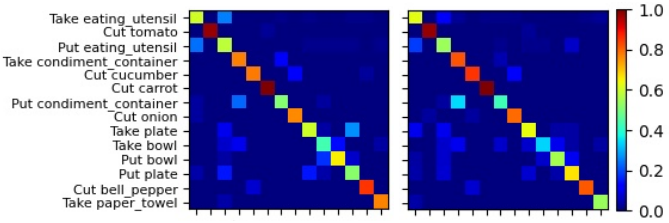
Object Size	Obfuscated?	Acc (Mean Class Acc)		
		Fine-grained Activity	Verb	Noun
Small	Yes	0.67 (0.63)	N/A	0.82 (0.82)
	No	0.67 (0.63)	N/A	0.81 (0.82)
Large	Yes	0.65 (0.57)	N/A	0.92 (0.94)
	No	0.79 (0.76)	N/A	0.94 (0.96)
All	Yes	0.63 (0.58)	0.74 (0.77)	0.79 (0.78)
	No	0.65 (0.62)	0.78 (0.79)	0.79 (0.8)

A. Impact of Object Size

Some objects (e.g., pot, pan, cutting board) are relatively larger than others (e.g., vegetables, small containers, knives) and so when the large objects appear in the scene, the hand is more occluded. According to the obfuscation function, defined in Section III, if no hand is present, the obfuscated image will be entirely the same color (a black frame) since there is no hand to assign the object. In such cases, the obfuscated image sequence will contain a few scattered black images, negatively impacting the model's ability to detect the activity. Figures 2a and 2b show the location distribution of the hands

Authorized licensed use limited to: Northwestern University. Downloaded on October 14,2022 at 19:51:33 UTC from IEEE Xplore. Restrictions apply.





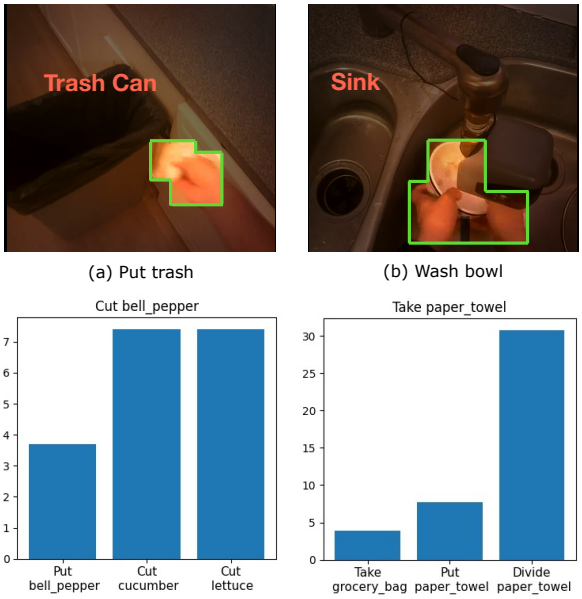


Fig. 4. (a)-(b): Classes where context might become important. Obfuscated images are insensible to the context in contrast to raw images. Green regions only show pixels that are visible on obfuscated image (c-d): Top 3 predicted classes for class "cut bell_pepper" and class "take paper_towel" when the raw image is used; The model confuses these classes with similar fine-grained activities.

and objects in-hand across images of a subject cutting a cucumber and using a cutting board, respectively. As shown with the cucumber, when the individual is using a smaller object such as a knife, the hand presence in the image is more dominant, while for activities involving large objects such as the cutting board, the hand is not visible most of the time.

B. Impact of Context Visibility

We noticed for some class labels that information from context, other than the hand and the object in-hand, is important. One good example is the class "Put trash" containing image sequences representing the moment when the wearer is throwing trash in a trash can. The model achieves an accuracy of 81% for this class when it is trained on raw image sequences. However, the accuracy drops to 70% when it is trained on obfuscated images. The potential reason behind this decline is that in the obfuscated image, only the *hand* and the *trash* in the hand are visible and, and while the object "trash" can refer to many things such as vegetable scraps, paper towel, etc., the important common feature that helps in recognizing this activity is the trash can. Another example where information in addition to the hand and object in-hand becomes important is the washing action, where the sink is important in providing useful information. Figures 4a and 4b present sample images for activity classes where context is important: putting the trash and washing a bowl.

C. Positive Impacts of Obfuscation

Surprisingly, for some classes, HOBM improves model accuracy substantially. In this section, we discuss the main reasons behind this improvement.

1) Presence of confounding objects: We noticed that too much unnecessary information on raw images can at times confuse the model. This happens when the model is exposed to a large set of objects present in the raw image. One example is where the subject is taking an object from the kitchen counter while other objects are placed nearby. We noticed often that in the raw image, the model confuses the object with nearby objects, whereas in the obfuscated image, the model is blind to confounding objects. Figures 2c and 2d show examples of this case where the model can predict the verb "take" successfully, but fails to predict the noun due to the presence of confounding objects.

2) Similar Fine-grained Activities: There exist a lot of similar classes in the dataset that makes the task of fine-grained activity recognition challenging. When the model is exposed only to the necessary information in the obfuscated image, it is forced to learn more robust features that are part of the activity class, and as a result performs better at distinguishing between similar activities. For example, the activity class "cut bell pepper" is very close to activities like "cut cucumber" or "cut lettuce." Likewise, "take paper towel" is similar to "put paper towel" and "divide/pull apart paper towel." We see that the raw-trained model often confuses these classes with each other, resulting in low class accuracy. Figures 4c and 4d show the classes most often confused with classes "cut bell pepper" and "take paper towel."

V. Conclusion

In this paper, we tested a mask obfuscation method's effect on hand-related human activity recognition accuracy. We show

Authorized licensed use limited to: Northwestern University. Downloaded on October 14,2022 at 19:51:33 UTC from IEEE Xplore. Restrictions apply.

on average a 2% reduction in accuracy when training and testing are done on obfuscated image sequences compared to the raw image sequences. We highlight the potential for obfuscation to maintain high utility in detecting fine-grained activities while addressing privacy concerns. This method further supports the principle of least privilege in that using less information does not necessarily reduce utility of data for machine learning methods. This method can streamline the process of capturing naturally occurring behavior, minimizing the amount of information captured and processed, and reduce the ethical burden of disseminating video datasets for the research community by addressing privacy concerns. Moreover, it can further assist researchers by potentially increasing the feasibility of wearing a camera in longitudinal studies.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation (NSF) under award number CNS1915847. We would also like to acknowledge support by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under award numbers K25DK113242 and R03DK127128, and National Institute of Biomedical Imaging and Bioengineering (NIBIB) under award number R21EB030305. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

REFERENCES

[1] J. Suto, S. Oniga, C. Lung, and I. Orha, "Comparison of offline and real-time human activity recognition results using machine learning techniques," *Neural Computing and Applications*, vol. 32, 10 2020.

[2] B. Bell, R. Alam, N. Alshurafa, E. Thomaz, A. Mondol, K. de la Haye, J. Stankovic, J. Lach, and D. Spruijt-Metz, "Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review," *npj Digital Medicine*, vol. 3, p. 38, 03 2020.

[3] A. Davies, V. Chan, A. Bauman, L. Signal, C. Hosking, L. Gemming, and M. Allman-Farinelli, "Using wearable cameras to monitor eating and drinking behaviours during transport journeys," *European Journal of Nutrition*, vol. 60, no. 4, pp. 1875–1885, Sep. 2020. [Online]. Available: https://doi.org/10.1007/s00394-020-02380-4

[4] M. Dimiccoli, J. Marín, and E. Thomaz, "Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, Jan. 2018. [Online]. Available: https://doi.org/10.1145/3161190

[5] A. Frome, G. K. M. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent, "Large-scale privacy protection in google street view," *2009 IEEE 12th International Conference on Computer Vision*, pp. 2373–2380, 2009.

[6] R. Alharbi, M. Tolba, L. C. Petito, J. Hester, and N. Alshurafa, "To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, sep 2019. [Online]. Available: https://doi.org/10.1145/3351230

[7] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2020. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2020.2986648

[8] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *2011 international conference on computer vision*. IEEE, 2011, pp. 407–414.

[9] M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 1–10. [Online]. Available: https://doi.org/10.1145/358916.358935

[10] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent, "Large-scale privacy protection in google street view," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2373–2380.

[11] J. Yan, F. Angelini, and S. M. Naqvi, "Image segmentation based privacy-preserving human action recognition for anomaly detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8931–8935.

[12] M. S. Ryoo, K. Kim, and H. J. Yang, "Extreme low resolution activity recognition with multi-siamese embedding learning," 2018.

[13] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," 2016.

[14] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia, "Enhancing lifelogging privacy by detecting screens," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 4309–4314. [Online]. Available: https://doi.org/10.1145/2858036.2858417

[15] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, "Recognition of activities of daily living with egocentric vision: A review," *Sensors*, vol. 16, no. 1, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/1/72

[16] Y. Iwasawa, K. Nakayama, I. Yairi, and Y. Matsuo, "Privacy issues regarding the application of dnns to activity-recognition using wearables and its countermeasures by use of adversarial training," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 1930–1936. [Online]. Available: https://doi.org/10.24963/ijcai.2017/268

[17] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Protecting sensory data against sensitive inferences," *CoRR*, vol. abs/1802.07802, 2018. [Online]. Available: http://arxiv.org/abs/1802.07802

[18] ——, "Mobile sensor data anonymization," in *Proceedings of the International Conference on Internet of Things Design and Implementation*, ser. IoTDI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 49–58. [Online]. Available: https://doi.org/10.1145/3302505.3310068

[19] S. A. Osia, A. Taheri, A. S. Shamsabadi, K. Katevas, H. Haddadi, and H. R. Rabiee, "Deep private-feature extraction," 2018.

[20] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," 2018.

[21] P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal, "Effects of degradations on deep neural network architectures," 2019.

[22] G. Kapidis, R. Poppe, E. van Dam, L. P. J. J. Noldus, and R. C. Veltkamp, "Egocentric hand track and object-based human action recognition," 2019.

[23] S. Sudhakaran, S. Escalera, and O. Lanz, "Lsta: Long short-term attention for egocentric action recognition," 2019.

[24] M. Schröder and H. Ritter, "Hand-object interaction detection with fully convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1236–1243.

[25] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik, "Reconstructing hand-object interactions in the wild," 2020.

[26] H. Fan, T. Zhuo, X. Yu, Y. Yang, and M. Kankanhalli, "Understanding atomic hand-object interaction with human intention," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[27] D. Shan, J. Geng, M. Shu, and D. Fouhey, "Understanding human hands in contact at internet scale," 2020.

[28] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *ECCV*, 2018.

[29] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[31] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017.

Authorized licensed use limited to: Northwestern University. Downloaded on October 14,2022 at 19:51:33 UTC from IEEE Xplore. Restrictions apply.