Robot Eye on Your Body Habits: Body-Focused Repetitive Behaviour Detection on Pepper

Zachary Syvenky, Samuel Antunes Miranda, Samir Arora, Salah Eddine Chahma, Zhitian Zhang, & Angelica Lim

Abstract— **Common repetitive behaviours like hair-pulling and nail-biting can be harmful when done frequently and subconsciously. These behaviours are known as Body Focused Repetitive Behaviours (BFRBs) and have been traditionally difficult to detect and treat due to their unconscious nature. Prior works mostly focus on using wearable devices to detect BFRBs. However, we propose to detect and interrupt harmful BFRBs using a humanoid robot - Pepper under a human-robot interaction scenario. We trained a convolutional neural network (CNN) using facial images captured from the Pepper robot to detect and interrupt BFRBs. Due to the lack of BFRBs data, we collected a dataset that consists of four BFRBs: nail-biting, hair-pulling, beard-pulling, eyebrow-pulling. We fine-tuned pre-trained CNN models using our collected dataset and show that our model could successfully detect these BFRBs. More specifically, our experimental results show that it was able to detect and interrupt nail-biting, as well as hair and eyebrow pulling with slightly lower precision. Beard pulling was more difficult to detect and our results did not show consistent detection. Our approach overall has demonstrated a higher precision than other existing methods of detecting these behaviours. The trained models and codebase can also be easily used in future research on detecting other forms of similar behaviours.**

I. INTRODUCTION

Many individuals engage in certain, mildly harmful behaviours such as hair pulling or nail biting during elevated periods of anxiety or boredom. These behaviours are known as Body Focused Repetitive Behaviours (BFRB). A study conducted by Houghton et al.[1] examined the self-reported prevalence of BFRBs among 4335 undergraduate students and found that 59.55% of the sample reported occasionally engaging in subclinical BFRBs and 12.27% met the criteria of pathological BFRB. The case of hair pulling has a specific diagnosis in the DSM-5 as trichotillomania [2], this is diagnosed in individuals where hair pulling is a chronic issue that negatively impacts their quality of life by producing bald spots or infections. This is not a rare condition, roughly 1% of the population has this diagnosis [3]. Nail biting (also known as onychophagy) can also cause issues such as infections or bleeding when it is done chronically.

Treating BFRBs is difficult, mainly because the behaviours are done subconsciously. Skurya et al. found that Habit Reversal Therapy (HRT) has seen the most success in treating BFRBs [4]. The authors found that HRT needs to be combined with mindfulness, i.e., increasing self-awareness and helping individuals recognize triggers for their BFRBs. However, recognizing that they are doing the behaviour is usually one of the hardest parts since these behaviours are typically performed subconsciously.

A common modern modality used in detecting BFRBs is wearable devices that use the position of the limb during an idle state, which is measured using Inertial Measurement Unit (IMU) devices. [5]. The primary issue with these devices is frequent false positives from similar motions that occur while doing harmless, everyday things. For example, using a screwdriver is similar to one of the actions of hair pulling. This means that the false positive rate for these devices can be quite high for the user and can quickly lead to frustration, and the abandonment of the device.

Our proposed solution to this problem is a real-time interactive software that runs on a Pepper robot[1] that can detect specific BFRBs with high precision. Once the robot detects these behaviours, it will interrupt and notify the user they are performing that behaviour.

Human Robot Interaction studies have shown several cases of robots successfully providing treatment for an array of conditions that typically require some degree of treatment such as autism [6], and dementia [7]. Prior research has shown that the use of robots to provide healthcare instructions is more effective and enjoyable for the user than prompts on a tablet [8] [9]. Humanoid robots have also been shown to be effective in implementing therapeutic treatment [10]. Therefore, a physical, humanoid robot providing the BFRB detection cues appears to be a promising approach.

Human-robot interaction in the area of BFRB treatment has been successfully done in a case study in the specific case of nail biting using a histogram of oriented gradients (HOG) method and a Cozmo robot [11]. Due to new advances in machine learning and computer vision algorithms, additional techniques such as CNNs are now more readily available, allowing us to design an approach that can detect several behaviours in the same model rather than just a single BFRB per model. As no existing public BFRB datasets exist from previous research, we created our own. It should be noted that we are not claiming to contribute a new technological advancement in machine learning or computer vision, but we rather report the results and contribute code for an application-specific system for behaviour detection of specific behaviours using a robot.

All authors are with the School of Computing Science, Simon Fraser University, Burnaby, BC, Canada. Corresponding author: zachary_syvenky@sfu.ca

[1] www.aldebaran.com/en/pepper

II. APPROACH

We divided this task into two sub-components: (i) detecting BFRBs (ii) taking action to interrupt. In the first part, the Pepper robot must detect if the subject is demonstrating a BRFB. For simplicity, we focus on three behaviours here: (a) hair pulling on the top of the scalp (referred to in this document as simply "hair pulling"), (b) facial hair pulling (composed of beard and eyebrow pulling), and (c) nail biting. These behaviours were chosen for two reasons. They are commonly occurring [1], and they are external behaviours that are performed around the face, which is the location the robot will focus on.

In the second part, the robot needs to correctly identify the behaviour, and audibly state that the user is performing such behaviour. For this project, the robot simply states that the user is performing the behaviour. It does not tell the person to stop doing the behaviour since that could be considered aggressive or irritating, and having the behaviour pointed out is enough for a rudimentary HRT treatment [4]. It does not perform any further therapeutic intervention, but that could be a potential area to expand upon in future research.

The BFRB behaviours we wish to capture are very specific and no public datasets currently exist, so we created a dataset ourselves. We created a data set of 5 different behaviours - nail-biting, hair-pulling, beard-pulling, eyebrow-pulling, and non-BFRB behaviours. During data collection, behaviours were acted due to the difficulties of capturing videos of naturalistic BFRBs at the scale of this project. The dataset consists of 9 people who took videos of themselves performing the BFRBs, as well as some non-BFRBs (e.g.: scratching their nose, brushing their hair to the side). They were then converted to a string of images at 5fps. These images were then categorized by the team members of the research team for use during model training. The dataset participants included 4 males and 5 females between the ages of 20 and 40 years old. The ethnicities of the actors were European, Moroccan, Brazilian, Iranian, Kyrgyz, Filipino, Chinese, and Indian.

At inference time, the Pepper robot was configured using the NaoQi library[2] to process the video stream from its primary camera. The images from that video stream were then passed into our model. If a behaviour was detected, the robot would highlight the behaviour. For example, when hair pulling is detected, the robot says, "You are pulling your hair".

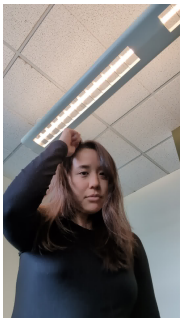
III. DATASET

The dataset consisted of 7088 image frames that were extracted from video recordings. We collected 1711 frames of hair-pulling, 663 frames of beard-pulling, 1616 frames of eyebrow-pulling, and 1616 frames of nail-biting. Non-BFRB movements such as yawning, putting hands on face, and standing still were also performed and 3099 frames were collected. A larger amount of non-BFRB actions (50% more than any single behavior) were recorded to reduce the number of false positives in the prediction given its frequent occurrence, as well as to demonstrate the real application scenario since most of the time, the subject will not be performing a BFRB. The dataset collection environment consisted of a variety conditions, such as variations in background, lighting, clothing, and camera angles in order to better represent real-world scenarios, as displayed in figures 1.a-e. The rationale for collecting and annotating a series of frames from videos over singular images is to better capture corner cases where the subject seems close to engaging in one of the behaviours, but is actually harmless, which may help generalization and reduce the number of false positives.

(a) Beard Pulling

(b) Non BFRB

(c) Nail Biting



(d) Hair Pulling

(e) Eyebrow Pulling

Fig. 1: Behaviour Examples

In order to train on a dataset of frames extracted from videos of 9 different individuals, we used a leave-one-out k-fold cross-validation method where k = 9. Models were trained on the labeled frames from 8 different people, and tested on the remaining individual. The validation set is constructed each fold by randomly assigning a person to each behaviour, and removing the images of that person performing that behaviour from the training set, and using it to validate instead.

To aid with generalization, data augmentation was implemented on the training dataset using transformations. The transformations chosen were horizontal flips, rotations by 10 degrees, and 20% zoom on the training images. In addition to creating training, validation, testing splits, and data augmentation, we normalized the pixel values by dividing each pixel value by 255, which helped make the training process more stable and gradients flow more smoothly during backpropagation. Moreover, the images were resized for each of the pre-trained models we used based on the specifications in their documentation. Full code is provided on the project website.[3]

[2]www.aldebaran.com

[3]https://github.com/ZachSyv/Pepper-BFRB-Detector

(a) Non BFRB (b) Hair Pulling

Fig. 2: Illustration of importance of fine-grained details: similar hand position can depict two different behaviours

IV. EXPERIMENTS

Following data collection and pre-processing, the objective was to identify a model that achieved the highest precision. To this end, we evaluated a range of pre-trained models (see Table I for more details). These pre-trained models leverage weights obtained through training on the ImageNet dataset, a large-scale collection of over 14 million hand-annotated images encompassing a diverse set of object categories.

Image processing versus skeleton processing. We inspected our dataset and selected an approach considering the challenging distinctions necessary in the classification tasks. For instance, in Figure 2, the hand position is the same in both the hair-pulling and non-BFRB example. The differentiating factor in the frames between hair pulling and non-BFRB is the presence of hair in the fingers in the image. Therefore, we reject an alternative approach using only skeleton tracking. Pixel-level image processing appears necessary to detect these behaviours.

Model architecture and fine-tuning. To capitalize on the knowledge embedded within these pre-trained models, we employed a transfer learning approach. Specifically, the convolutional layers of a pre-trained model were frozen, preserving their learned features. A 2D convolutional layer with 64 filters with a ReLU non-linearity activation function was then added on top of the pre-trained layers. A single dense layer with 256 units, ReLU non-linearity, and L2 regularization was then added. To mitigate overfitting, 2D global average pooling was used after the 2D convolutional layers, and a Dropout layer with a rate of 0.5 was incorporated after the dense later. Finally, a single output layer with 5 units and softmax activation was appended, aligning with the multi-class classification task. The training process employed a batch size of 16 for both the training and validation datasets. A learning rate of 0.0001 and a categorical cross-entropy loss function was utilized, reflecting the multi-class nature of the problem. Each model was trained for 10 epochs.

An ablation study showing the result without the 2D convolutional layer was conducted, and the results are shown in Figure 3 in parentheses.

Accuracy and speed trade-offs. As evident in Table I, InceptionResNetV2 performed the best overall. However, its large number of parameters limit its practicality in the resource-constrained scenarios of running the model off of Pepper. The Xception model emerged as a compelling choice. It possessed the near half the number of parameters as InceptionResNetV2, and had a similar F1 score. Despite VGG being even smaller, the small amount of trainable parameters seems to have harmed the F1 score.

TABLE I: Comparison of models on video stream test

Model	# Params	# Params trainable	Precision	F1-score
Xception	22.06m	1.20m	0.49	0.47
VGG19	20.34m	312.90k	0.5	0.35
InceptionResNetV2	55.24m	902.73k	0.59	0.51

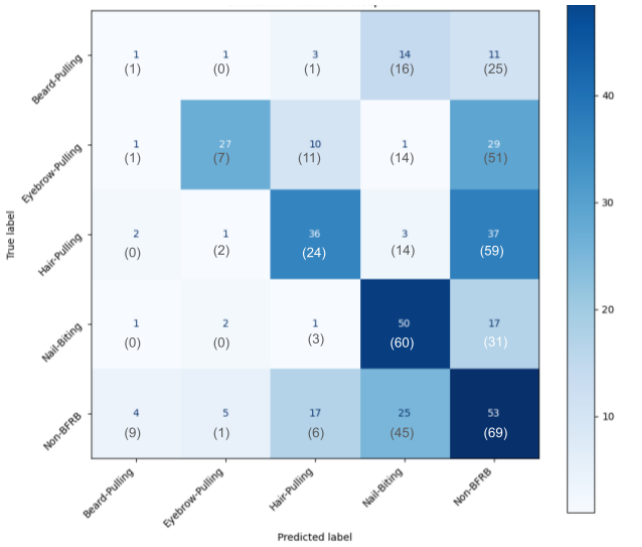


Fig. 3: Confusion matrix of video stream results for the Xception model on the test set. Numbers in parentheses are the results from the ablation study on the Xception model without the added CNN layer

Video stream processing. In order to evaluate the performance of the implementation when deployed on Pepper, test were done by passing a sequence of 25 frames from the original videos to the model. This number is to replicate 5 seconds with the 5fps camera on Pepper. If, during that time frame, 3 of the same behaviour is detected with a confidence of > 0.45, except for nail biting which we set to > 0.70, then it would classify that behaviour. The confidence thresholds were decided through trial and error.

V. RESULTS

Figure 3 & Table II show the combined results across all folds for the video stream processing experiment for the Xception architecture. As we can observe, nail-biting behaviours were detected the best during testing. No nail-

TABLE II: Classification results for Xception video stream test results. Numbers in parentheses are the results from the ablation study on the model without the added CNN layer

Behaviour	Precision	Recall	F1-Score
Beard-Pulling	0.11 (0.10)	0.03 (0.02)	0.05 (0.04)
Eyebrow-Pulling	0.75 (0.70)	0.40 (0.09)	0.52 (0.15)
Hair-Pulling	0.54 (0.55)	0.46 (0.24)	0.49 (0.34)
Nail-Biting	0.54 (0.40)	0.70 (0.64)	0.61 (0.49)
Non-BFRB	0.36 (0.29)	0.51 (0.53)	0.42 (0.38)

(a) (b) (c)

(d) (e)

Fig. 4: Misclassification Examples. a: non-BFRB as beard pulling; b: non-BFRB as hair pulling; c: eyebrow pulling as hair pulling; d & e: non-BFRB as nail biting

biting videos were typically given an incorrect label besides non-BFRB. Nail biting was also the most common false label to be applied besides non-BFRB. Several non-BFRB actions that took place near the mouth were classified as nail-biting. Beard pulling in particular was most commonly mistaken as nail-biting, followed by non-BFRB. Hair pulling was also detected quite well. Some cases where the video subject had long hair were misclassified as nail biting, but many were classified correctly. Sometimes the strands of hair being pulled were difficult to see and were misclassified as non-BFRB. Many non-BFRB actions that took place near the top of the head were misclassified as hair-pulling. Eyebrow pulling detection also performed well but in some cases was detected as hair pulling, specifically during instances where the subjects were pulling the side of their eyebrow by their hairline. Beard-pulling detection was quite poor. The behaviour was usually labeled as nail-biting or non-BFRB. False positives were typically made for nail biting and hair pulling when the action was close to the mouth or scalp respectively. Misclassification examples can be found in Figure 4. These examples highlight the ambiguity of many of these actions in our frame-based paradigm. Specifically, 3c was misclassified as hair pulling, and 3b running their hand by their hair was misclassified as hair-pulling.

We implemented the system on Pepper and a video demonstration is available. [4]

VI. DISCUSSION AND LIMITATIONS

The F1 score of each behaviour may relate to how similar the characteristics of the behaviours are between the 9 participants' data. As the participants have similar teeth and hands, nail-biting was very similar among them, and was detected well. Eyebrows were also similar, but sometimes obstructed by hair or glasses so detection was more difficult. Hair length, style, and colour differed, so hair pulling was slightly less accurate. The beard-pulling behaviour was particularly hard to distinguish from other harmless behaviours, like resting a hand on the chin when the camera was much lower than the face, and was often not detected as anything. Beard pulling also had less data collected due to 5 of the participants having no beard hair, and those that did have beards varied significantly in the prominence of the beard, leading to the behaviour looking different among each individual, making detection difficult.

Other limitations included a small sample size resulting in a limited diversity in participants; for example, we did not have any participants with dark skin colour. Our dataset also did not contain samples with characteristics like light blond hair, bald or shaved head, brightly dyed hair, or piercings. Therefore, the effectiveness of our model on people who have these characteristics is unknown.

A. Future Work

This was a proof-of-concept study. As our results suggest that this approach can work in some capacity, there are many potential areas to expand upon. A full-scale study and trial for this proposed method and model on the Pepper robot would be the next step. Improvements like a larger sample size could be made, and rigorous implementation tests could be performed. Having the model itself work on a series of images instead of a single one would also be a worthwhile approach. This would be able to help performance metrics like accuracy and precision, reduce latency time in the robot's response, and allow more intricate behaviours to be detected that are hard to capture with just a single frame, for example, chewing the inside of one's cheek. Implementing techniques such as LSTM may also help in the same domain. There are also several other BFRBs that could be detected like eyelash picking, chronic eye rubbing, picking skin around lips, or scab picking. Even behaviours that are harmless, but may be undesirable habits like nose-picking could be detected. Finally, expanding on the robot's ability to provide therapeutic intervention after detection would be a very promising area. However, more research is needed to determine what type of therapeutic intervention should be implemented.

VII. CONCLUSIONS

We introduced a technique to detect and interrupt behaviours, like nail biting and hair/facial hair pulling, that are typically difficult to detect due to their unconscious nature, with the goal of minimizing false positive detections. This was done by creating a dataset with labelled images of these behaviours, fine-tuning an existing CNN to detect these behaviours, and then running the model to interact with a Pepper robot, which, upon detection, could interrupt these behaviours. This study shows promising results in the precision of detection and can likely be further improved in the future to be reliable for detection and intervention.

[4] https://www.youtube.com/watch?v=3ig3vNkJBN0

REFERENCES

[1] Houghton, D. C., Alexander, J. R., Bauer, C. C., & Woods, D. W. (2018). Body-focused repetitive behaviors: More prevalent than once thought? Psychiatry Research, 270, 389–393. https://doi.org/10.1016/j.psychres.2018.10.002

[2] American Psychiatric Association, Diagnostic and statistical manual of mental disorders, 5th ed. Boston: Pearson, 2013.

[3] G. J. Diefenbach, D. Reitman, and D. A. Williamson, "Trichotillomania: A challenge to research and practice ," Clinical Psychology Review, vol. 20, no. 3, pp. 289–309, Apr. 2000, doi: https://doi.org/10.1016/s0272-7358(98)00083-x.

[4] Skurya, J., Jafferany, M., & Everett, G. J. (2020). Habit reversal therapy in the management of body focused repetitive behavior disorders. Dermatologic Therapy, 33(6). https://doi.org/10.1111/dth.13811

[5] Benjamin Lucas Searle, Dimitris Spathis, M. Constantinides, D. Quercia, and C. Mascolo, "Anticipatory Detection of Compulsive Body-focused Repetitive Behaviors with Wearables," arXiv (Cornell University), Sep. 2021, doi: https://doi.org/10.1145/3447526.3472061.

[6] E. Kim, R. Paul, F. Shic, and B. Scassellati, "Bridging the Research Gap: Making HRI Useful to Individuals with Autism," Journal of Human-Robot Interaction, pp. 26–54, Aug. 2012, doi: https://doi.org/10.5898/jhri.1.1.kim.

[7] F. Martín Rico et al., "Robots in therapy for dementia patients," rua.ua.es, Jan. 2013, doi: https://doi.org/10.14198/JoPha.2013.7.1.07.

[8] J. A. Mann, B. A. MacDonald, I.-H. Kuo, X. Li, and E. Broadbent, "People respond better to robots than computer tablets delivering healthcare instructions," Computers in Human Behavior, vol. 43, pp. 112–117, Feb. 2015, doi: https://doi.org/10.1016/j.chb.2014.10.029.

[9] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," International Journal of Human-Computer Studies, vol. 77, pp. 23–37, May 2015, doi: https://doi.org/10.1016/j.ijhcs.2015.01.001.

[10] J. Camargo and A. Young, "Combined Strategy of Machine Vision with a Robotic Assistant for Nail Biting Prevention," 2017 14th Conference on Computer and Robot Vision (CRV), Edmonton, AB, Canada, 2017, pp. 205-208, doi: 10.1109/CRV.2017.57.

[11] T. Nomura, "Consideration of Mental Therapeutic Robots from Psychological and Sociological Perspectives," SCIS & ISIS SCIS & ISIS 2006, vol. 2006, pp. 1476–1480, Jan. 2006, doi: https://doi.org/10.14864/softscis.2006.0.1476.0.