Providing Hand Use Context for Outpatient Neurorehabilitation with Egocentric Object Detection

Adesh Kadambi[†‡] and José Zariffa[†‡]
[†]*KITE - Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada*
[‡]*Institute of Biomedical Engineering, University of Toronto, Toronto, ON, Canada*
adesh.kadambi@mail.utoronto.ca, jose.zariffa@utoronto.ca

***Abstract*—Advancements in wearable technology and machine learning have the potential to enhance rehabilitation therapy planning, particularly in outpatient settings by capturing context-specific information about an individual's hand use, including interactions and activities of daily living (ADLs). In this study, we evaluated the performance of two object detection models, Detic and UniDet, on egocentric videos recorded by individuals with spinal cord injury (SCI). Our evaluations revealed that UniDet, when evaluated on its original 700 classes, achieved a Mean Average Precision (mAP) of 0.038 for all objects and 0.099 for active objects. When evaluated on a set of 27 consolidated functional categories, UniDet's performance improved to an mAP of 0.16 for all objects and 0.19 for active objects. Detic demonstrated superior performance with an mAP of 0.19 for all objects and 0.30 for active objects when evaluated on the 27 functional categories. However, the ground truth labelling strategy resulted in a large number of false positives, suggesting that the model performance is likely higher. Despite challenges posed by low-light conditions and motion blur, this study provides crucial insights into the potential of object detection models in therapy planning, facilitating the integration of wearable technology and machine learning in outpatient rehabilitation and enabling more personalized and effective therapeutic strategies.**

***Clinical relevance*— The ability to encode context from egocentric videos of patients' daily activities presents a transformative opportunity in outpatient neurorehabilitation, enabling clinicians to develop more personalized and effective rehabilitation strategies grounded in real-world hand usage patterns.**

***Index Terms*—egocentric video, object detection, spinal cord injury, neurorehabilitation, wearable technology**

I. Introduction

Stroke and spinal cord injuries (SCI) present significant challenges to the ability of affected individuals to live independently and perform activities of daily living (ADLs). These conditions can severely impair motor function, requiring tailored rehabilitation approaches to regain independence as much as possible [1]. In this context, occupational and physical therapists often assess ADLs and in-clinic capacity [2] (i.e., what an individual can do in a standardized environment) to guide outpatient rehabilitation and therapy planning.

This approach assumes that any improvements in in-clinic capacity will translate to improvements in performance [2] (i.e., what an individual actually does in their daily environment). Recent studies, however, indicate that improvements in capacity do not always correspond to changes in performance, suggesting that more comprehensive methods for assessing real-world performance are needed [3].

Traditional methods of gathering context-specific information about ADLs, such as direct observation or self-reporting, have several drawbacks, including self-report bias and the inability to mimic individual home environments [4]. Wearable technologies, however, can mitigate these issues by opening new paths for data collection [5]. Specifically, egocentric videos offer a first-person view of ADLs, capturing extensive human-centric data in naturalistic settings. This provides rich insights into the activities, interactions, and strategies of individuals with disabilities, surpassing other wearable sensors like accelerometers and magnetic sensors in context [6]–[9].

Leveraging this technology, we previously developed a dashboard that reports hand performance measures for outpatient rehabilitation using head-mounted egocentric cameras [10], [11]. Clinicians acknowledged its potential to monitor rehabilitation progress, deliver feedback on hand use, and track improvement over time. However, they have emphasized the need for more nuanced contextual information, such as object interactions or activities being performed, for meaningful interpretation of metrics or effective therapy planning.

With the rapid advancement of object detection models in the field of computer vision, new opportunities are emerging to automatically detect and classify objects within egocentric video footage, thereby encoding context in a scalable manner from object interactions. We aimed to evaluate the efficacy of models (e.g., Detic [12] and UniDet [13]) in detecting various object classes in environments commonly found in home-based settings, focusing on egocentric videos recorded by individuals with stroke and SCI during outpatient therapy. Our assessment covers their accuracy in detecting and classifying objects 'in the wild' and explores their appropriateness for supplying the contextual information vital for informed therapy planning, based on the premise that object interactions correlate directly with ADLs.

By considering both the quantitative performance and the nature of the errors made by these models, our work offers a comprehensive insight into their potential application and

This work was supported by the Craig H. Neilsen Foundation (grant number 542675), the Praxis Spinal Cord Institute, and the Ontario Early Researcher Award program (ER16–12-013).

limitations in outpatient rehabilitation settings. This understanding of object detection model performance in a rehabilitation framework contributes to the broader discourse on the integration of wearable and machine learning technologies in therapeutic practices for individuals with conditions such as stroke or SCI.

II. Methods

A. Dataset

We used a dataset comprised of 2261 minutes of egocentric video recordings derived from 16 participants, all of whom demonstrated impaired hand functionality as a consequence of SCI. This footage was procured from [11], which adhered to the recording protocol delineated by [14]. The participants, involved in an array of activities of daily living (ADLs) within their natural environments, were recorded without any imposed constraints. The duration of recorded footage varied per participant, ranging from a minimum of 7 minutes to a maximum of 229 minutes, with an average duration of 141.31 minutes (sd = 72.91 minutes).

The original recordings were segmented into 1-minute snippets and classified into one of 7 predefined ADL categories based on the participant's actions observed within the snippet: Communication Management (428 instances), Functional Mobility (207 instances), Grooming & Health Management (172 instances), Home Management (407 instances), Meal Preparation and Cleanup (625 instances), Self Feeding (257 instances), and Leisure & Other Activities (165 instances). Snippets containing sensitive information or devoid of object interactions or hand movements were excluded from the dataset. In cases where multiple ADLs were observed in a single snippet, the snippet was assigned the label of the predominant ADL (i.e., the one performed for the longest duration within the minute).

A stratified sampling approach was implemented, wherein two videos from each participant per ADL category were randomly selected. Frames were then extracted at a rate of 1 FPS from these selected videos. Object annotations were made in the extracted frames, wherein bounding boxes were drawn around objects pertinent to the ADL being performed and subsequently labelled using the 700 classes in the unified label space from [13]. These labels were further consolidated into 27 distinct object classes based on their functional similarities: animal, food, plant, sports equipment, wheelchair or walker, home appliance or tool, kitchen or cooking utensil, tableware, drinkware, kitchen appliance, furniture, cabinetry, furnishing, house fixture, electronics, tv or computer, phone or tablet, cleaning product, toiletry, bathroom fixture, sink, office stationary, clothing, footwear, clothing accessory, bag, and other. The processed data used for model evaluation is comprised of 1482 images, hosting a cumulative total of 4757 object annotations.

B. Model Selection and Evaluation

For our study, the critical criterion in model selection was the ability to detect a multitude of common household objects since these models performed inference on videos from the home environments of individuals with stroke or spinal cord injury. Two models, Detic [12] and UniDet [13], were selected for evaluation as representative examples that met this criterion.

The models were evaluated quantitatively using the Mean Average Precision (mAP) metric, focusing on two distinct tasks: detecting all objects and detecting 'active objects'—those objects in contact with the participant's hands, identified using the Egocentric 100DOH model proposed in [15]. Additionally, we explored the change in performance obtained by grouping functionally similar objects, reflecting our study's emphasis on the contextual understanding of rehabilitation settings.

A qualitative analysis was conducted to deepen our understanding of model performance by identifying common misclassification themes and uncovering potential reasons underlying model failures. This multifaceted evaluation approach highlighted the models' capabilities and limitations in detecting household objects within a rehabilitation context, in line with our study's primary objective.

III. Results

A. Quantitative Evaluation

The UniDet model was initially evaluated on all 700 original classes, attaining mAP@0.5 scores of 0.038 for all objects and 0.099 for active objects. Its performance improved significantly when consolidated to 27 object classes, achieving scores of 0.16 and 0.19 for all and active objects, respectively. Meanwhile, the Detic model, evaluated only on these 27 classes (due to the dataset's original labelling with the 700 Unidet classes), outperformed UniDet with mAP@0.5 scores of 0.19 and 0.30 for all and active objects. A detailed comparison of the two models' performance is provided in Table I.

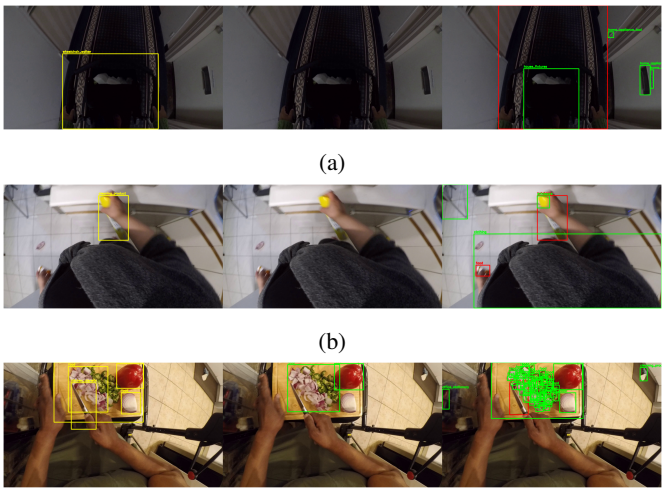
B. Qualitative Evaluation

Qualitative analysis revealed a few key challenges that potentially impacted the performance of the models. First, the presence of dark images in the dataset hindered the models' ability to accurately detect objects (Fig. 1a). Second, motion blur, a common occurrence in egocentric videos, posed a significant challenge to accurate object detection (Fig. 1b).

The Detic model often detected many objects that were not labelled in the ground truth, with a considerable portion of detections being accurate (Fig. 1c). This is further supported by observing the distribution of labels and prediction (Fig. 2).

TABLE I: Comparative evaluation of detection models

Model	Objects	Classes	mAP@0.5
Unidet	All	700	0.038
Unidet	Active	700	0.099
Unidet	All	27	0.16
Unidet	Active	27	0.19
Detic	All	27	**0.19**
Detic	Active	27	**0.30**



(a)

(b)

(c)

Fig. 1: Examples of images where models did not detect any objects due to (a) low-light images and (b) motion blur; (c) there were also instances where the Detic model in particular, would accurately detect objects that were not labelled in the ground truth which could negatively affect model performance. In all subfigures, the ground truth boxes are on the left, Unidet predictions are in the middle, and Detic predictions are on the right. Red bounding boxes denote predicted active objects.

Since the mAP metric is a measure of the precision and recall of the model, the detection of these unlabeled objects would be considered as false positives, thereby decreasing the precision and consequently the mAP. This suggests that the mAP scores reported might underestimate the models' actual performance in terms of object detection.

These results provide important insights into the capabilities and limitations of the evaluated models in the context of rehabilitation. The Detic model overall exhibited superior performance, however, both models were affected by challenging conditions such as dark images and motion blur. Furthermore, the detection of unlabelled objects by the models suggests potential areas for improvement in the ground truth labelling process, which could in turn improve the models' evaluated performance.

IV. Discussion

In object detection for therapy planning, we should focus on the functional use of objects rather than their visual similarity due to the importance of discerning ADLs, which serve as the foundational benchmarks in gauging the efficacy of hand use during the outpatient rehabilitation process.

Traditional object detection frameworks tend to cluster objects based on shared visual characteristics, an approach that may not align optimally with the objectives of rehabilitation where the functional role of an object, in relation to the specific activity being performed, carries a higher significance than its visual attributes. For example, a knife and a pen may bear a visual resemblance but are used in vastly different manners during their respective ADLs (i.e., self-feeding vs. communication management). Hence, shifting towards consolidating categories based on function instead of appearance may enhance the model's applicability to a rehabilitation context.

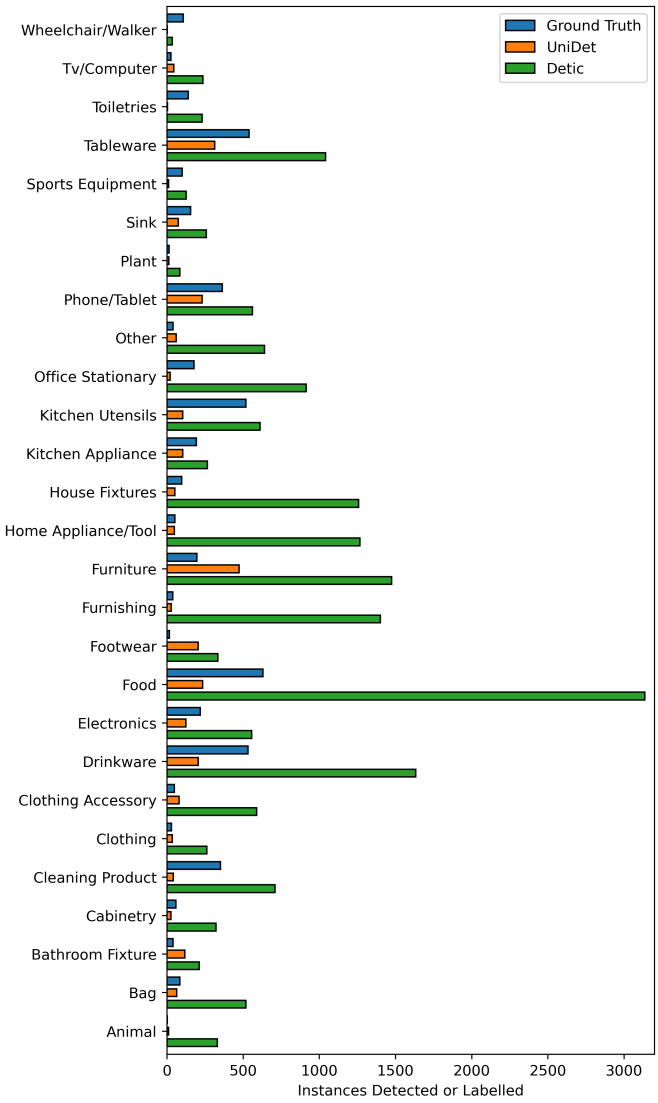


Fig. 2: Distribution of predicted versus ground truth labels.

We used the original models, which were trained on 700 classes for Unidet and 1203 classes for Detic, without any retraining. These labels were then manually grouped into 27 functional categories, and the evaluation was performed on these consolidated categories. This substantial reduction in classes led to a notable rise in model performance (from 0.038 to 0.16 mAP), improved the user-friendliness of delivering insights to clinicians, and potentially delivered more contextually pertinent information to clinicians since functional object categories are tightly coupled with specific ADLs being performed in the videos.

Low-light conditions and motion blur notably hindered model performance. Low-light environments, characterized by high noise, bad illumination, and low contrast, made feature

detection difficult. Motion blur, caused by head or hand movements, could obscure object boundaries and create spatial confusion, impacting object detection. Strategies to mitigate these include image enhancement algorithms for low-light situations and custom label generation for motion blur [16], [17]. Additionally, attention mechanisms can help models focus on areas more likely to contain objects relevant to ADLs being performed in videos, enhancing detection efficiency and reliability in outpatient rehabilitation.

While the results are promising, there are several limitations to consider. For instance, ground truth labels were only assigned to objects that were relevant to the ADL being performed (e.g., if the participant was self-feeding, a plate and spoon would be labelled, but the notebook next to them would not be labelled), which could result in potential discrepancies in model evaluation. As mentioned previously, if a model correctly identifies an object that was not included in the ground truth labels due to its perceived irrelevance to the ADL by the labeler, it could artificially reduce performance metrics. Future studies could address this limitation by employing a more comprehensive labelling strategy that includes all objects present, regardless of their direct relevance to the ADL, thereby giving a more accurate representation of model performance. Nonetheless, the analysis focusing on active objects in this study partially mitigates this limitation.

Furthermore, the ground truth was labelled using UniDet's 700 class label space, allowing for a comparison between the performance of UniDet's original classes and the consolidated 27 functional categories. However, a similar evaluation could not be conducted for the Detic model due to differences in classification schemes. This disparity presents a limitation in the direct comparison of the two models' performances.

Another limitation lies in the scope of the data used. The current study was conducted using a limited dataset, which may not fully represent the diversity of real-world scenarios and ADLs. The dynamic nature of ADLs and the inherent variability in how different individuals perform the same tasks present a unique challenge to model performance. Future work may need to account for this variability by incorporating individual-specific training data or developing better methods to handle task variability.

Other avenues for inquiry include exploring different label consolidation strategies, such as grouping by functional category, as done in this study, versus grouping by visual similarity. Through this, a more nuanced understanding of the effects of label consolidation strategies on model performance and interpretability could be developed.

V. Conclusion

This study explored the potential of object detection models to recognize a multitude of common household objects in egocentric videos recorded by individuals with stroke or SCI and provide therapists with enriched contextual information about hand use at home, thereby enabling more tailored therapy planning. Results demonstrated the potential to provide meaningful contextual information for rehabilitation therapy planning using object detection. Consolidating object categories based on functional use significantly improved the models' performance and increased their applicability in rehabilitation. However, this research also highlights challenges, such as low-light conditions and motion blur, that hindered the models' effectiveness. These findings emphasize the need for specialized handling of such issues and an enhanced ground truth labelling strategy. Overall, this work underscores the potential of integrating machine learning models with wearable technology for outpatient neurorehabilitation. With continual refinement, these models can facilitate personalized therapeutic strategies based on real-world hand usage patterns, leading to more targeted treatment and improved patient outcomes.

References

[1] K. Phillips, "Maximising independence through community rehabilitation," *International Journal of Disability Management*, vol. 9, 2014.

[2] World Health Organization, "International classification of functioning, disability and health (ICF)," 2018, accessed: 2023-1-9.

[3] K. J. Waddell, M. J. Strube, R. R. Bailey, J. W. Klaesner, R. L. Birkenmeier, A. W. Dromerick, and C. E. Lang, "Does Task-Specific training improve upper limb performance in daily life poststroke?" *Neurorehabil. Neural Repair*, vol. 31, no. 3, pp. 290–300, Mar. 2017.

[4] A. S. Adams, S. B. Soumerai, J. Lomas, and D. Ross-Degnan, "Evidence of self-report bias in assessing adherence to guidelines," *Int. J. Qual. Health Care*, vol. 11, no. 3, pp. 187–192, Jun. 1999.

[5] C. Adans-Dester, N. Hankov, A. O'Brien, G. Vergara-Diaz, R. Black-Schaffer, R. Zafonte, J. Dy, S. I. Lee, and P. Bonato, "Enabling precision rehabilitation interventions using wearable sensors and machine learning to track motor recovery," *NPJ Digit Med*, vol. 3, p. 121, Sep. 2020.

[6] M. Noorkõiv, H. Rodgers, and C. I. Price, "Accelerometer measurement of upper extremity movement after stroke: a systematic review of clinical studies," *J. Neuroeng. Rehabil.*, vol. 11, p. 144, Oct. 2014.

[7] N. Friedman, J. B. Rowe, D. J. Reinkensmeyer, and M. Bachman, "The manumeter: a wearable device for monitoring daily use of the wrist and fingers," *IEEE J Biomed Health Inform*, vol. 18, no. 6, pp. 1804–1812, Nov. 2014.

[8] N. P. Oess, J. Wanek, and A. Curt, "Design and evaluation of a low-cost instrumented glove for hand function assessment," *J. Neuroeng. Rehabil.*, vol. 9, p. 2, Jan. 2012.

[9] S. I. Lee, X. Liu, S. Rajan, N. Ramasarma, E. K. Choe, and P. Bonato, "A novel upper-limb function measure derived from finger-worn sensor data collected in a free-living setting," *PLoS One*, vol. 14, no. 3, p. e0212484, Mar. 2019.

[10] A. Bandini, A. Kadambi, R. D. Ramkalawan, S. L. Hitzig, and J. Zariffa, "A web-based interface for monitoring hand use in people with cervical spinal cord injury living in the community," *Journal of Spinal Cord Medicine*, vol. 44, no. SUPPL1, p. S320, Sep. 2021.

[11] A. Bandini, M. Dousty, S. L. Hitzig, B. C. Craven, S. Kalsi-Ryan, and J. Zariffa, "Measuring hand use in the home after cervical spinal cord injury using egocentric video," *J. Neurotrauma*, Jul. 2022.

[12] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting Twenty-Thousand classes using Image-Level supervision," in *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 350–368.

[13] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," Feb. 2021.

[14] M.-F. Tsai, A. Bandini, R. H. Wang, and J. Zariffa, "Capturing representative hand use at home using egocentric video in individuals with upper limb impairment," *J. Vis. Exp.*, no. 166, Dec. 2020.

[15] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.

[16] W. Chen and T. Shah, "Exploring low-light object detection techniques," Jul. 2021.

[17] M. Sayed and G. Brostow, "Improved handling of motion blur in online object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 1706–1716.