Detecting Activities of Daily Living in Egocentric Video to Contextualize Hand Use at Home in Outpatient Neurorehabilitation Settings

Adesh Kadambi, *Member, IEEE* and José Zariffa, *Senior Member, IEEE*

***Abstract*— Wearable egocentric cameras and machine learning has the potential to provide clinicians with a more nuanced understanding of patient hand use at home after stroke and spinal cord injury (SCI). However, they require detailed contextual information (i.e., activities and object interactions) to effectively interpret metrics and meaningfully guide therapy planning. In this study, we propose an object-centric approach for Activities of Daily Living (ADL) recognition that employs pre-trained object detection and hand-object interaction modules to generate a bag of objects, which serves as the primary indicator of activities. We evaluated our models on a complex dataset collected in the wild comprising of 2261 minutes of egocentric video from 16 participants with impaired hand functionality using leave-one-subject-out cross validation and reported the mean F1-score across participants and the percentage of participants who had and F1-score greater than 0.5. Our best performing model, a logistic regression using object binary presence and active objects as input features, achieved a mean weighted F1-score of 0.78 $\pm$ 0.12, with 100% of participants having an F1-score greater than 0.5. Our framework demonstrates that object information is sufficient for accurately classifying ADLs in a rehabilitation context that requires robustness to variable environments, activities, and hand impairments.**

***Index Terms*— activity detection, egocentric video, hand-object interaction, object detection, outpatient neurorehabilitation, spinal cord injury, stroke, wearable technology**

I. Introduction

RECOVERING hand function, and in turn, the ability to perform Activities of Daily Living (ADLs) independently, is a top priority for individuals with stroke or spinal cord injury (SCI) once they return to the community [1], [2]. These activities, ranging from basic personal care to complex tasks requiring fine motor skills, are critical for the affected individuals' autonomy, quality of life, and integration into society. Traditionally, occupational and physical therapists have relied on direct observation and patient self-reporting to assess hand use and the ability to perform ADLs. However, these methods are limited by biases in self-reporting and the inability to capture the diversity of home environments [3]. Therefore, a more nuanced understanding of hand function in naturalistic settings is crucial for designing effective rehabilitation strategies that align with patients' real-world needs.

To this end, we previously developed an algorithmic framework [4] employing wearable egocentric cameras to document hand use at home, delivering performance metrics to clinicians via a dashboard [5]. While clinicians recognized the potential of the dashboard for monitoring rehabilitation and providing feedback to their patients, they have expressed the need for more detailed contextual information—such as the breakdown of activities and object usage—to effectively interpret metrics and meaningfully guide therapy planning [5].

Recent advancements in egocentric activity recognition have been driven by the availability of extensive egocentric datasets like EPIC-KITCHENS [6], [7] and Ego4D [8]. These advancements include the use of convolutional neural networks (CNNs) for activity recognition. 2D-CNNs can extract spatial features at the frame level and aggregate these temporally via average pooling [9], recurrent neural networks [10], or Temporal Shift Modules [11]. Alternatively, 3D-CNNs extend 2D-CNNs to incorporate the temporal dimension directly into the convolutional process, allowing for a richer understanding of video context and motion [12], [13]. Despite their higher computational demands, 3D-CNNs provide robust performance, as exemplified by the SlowFast network [14], which processes videos at varying frame rates to simultaneously capture both dynamic and nuanced aspects of human actions, making it particularly suitable for analyzing complex activities captured in egocentric videos.

More recently, transformers [15], although originally developed for NLP tasks, have been adapted for video-based activities, offering advantages over CNNs by globally analyzing data relationships through attention mechanisms. Notable implementations, such as the Video Swin Transformer [16] and TimeSformer [17], illustrate the efficacy of transformers in handling complex video understanding tasks like activity recognition.

While these methods have demonstrated their effectiveness

This work was supported by the Craig H. Neilsen Foundation (grant number 542675), the Praxis Spinal Cord Institute, the Ontario Early Researcher Award program (ER16–12-013), and the Canadian Institutes of Health Research (grant number 13556838).

Adesh Kadambi is with the KITE - Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada and the Institute of Biomedical Engineering, University of Toronto, Toronto, ON, Canada (e-mail: adesh.kadambi@mail.utoronto.ca).

José Zariffa is with the KITE - Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada, and the Institute of Biomedical Engineering, the Rehabilitation Sciences Institute, Faculty of Medicine, and the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada (e-mail: jose.zariffa@utoronto.ca).

in various video understanding tasks, our objective prioritizes practical application "in the wild". Therefore, we adopt an object-centric approach, leveraging both active and passive objects to infer ADLs. This methodology focuses on object interaction—a crucial aspect of activity recognition which has been previously validated in literature [18]–[22]—and is particularly beneficial for our target demographic since it abstracts away the variability in how activities are performed due to compensatory movements, which are common in individuals with upper-limb impairment. This ensures our solution is not just technically proficient but is also practically meaningful by fulfilling therapists' informational needs regarding object interactions and ADLs performed. The primary contribution of our study is to demonstrate that object information is sufficient to achieve ADL recognition in a manner that is robust to variable environments, activities, and hand impairments, as demonstrated on a complex dataset collected in the wild.

II. Methods

A. Dataset

We used a dataset comprising 2261 minutes of egocentric video recordings obtained from 16 participants with impaired hand functionality due to SCI. The collection of this dataset was previously described by Bandini *et al.* [4] and followed the recording protocol outlined by Tsai *et al.* [23]. The participants, involved in an array of real ADLs within their home 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 ± 72.91 minutes.

The recordings were segmented into 1-minute snippets and classified into seven predefined ADL categories chosen based on the most common activities observed in the dataset: 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).

B. Definitions

The chosen ADL categories align with the American Occupational Therapy Association's Occupational Therapy Practice Framework [24]. These categories represent a spectrum of functional activities commonly encountered by individuals with impaired hand functionality:

- **Self-Feeding:** Involves tasks related to setting up, arranging, and bringing food or fluid from the plate or cup to the mouth.
- **Functional Mobility:** Encompasses movements from one position or place to another, including in-bed mobility, wheelchair mobility, transfers, and functional ambulation.
- **Grooming & Health Management:** Includes activities such as obtaining and using supplies, personal grooming (e.g., haircare, skincare), and managing health and wellness routines (e.g., exercise, medication routines).
- **Communication Management:** Covers the exchange and interpretation of information using various communication tools and equipment (e.g., phone, laptop).
- **Home Establishment and Management:** Involves tasks related to maintaining personal and household possessions and environments.
- **Meal Preparation and Cleanup:** Encompasses planning, preparing, serving meals, and cleaning up afterward.
- **Leisure & Other Activities:** Represents non-obligatory activities engaged in during discretionary time.

C. Feature Engineering

The process for generating features for ADL classification is depicted in Fig. 1. Frames were extracted at a rate of 1 FPS from each 1-minute video snippet. The Detic object detection model [25] provided information on object presence, and the 100DOH hand-object interaction model [26] was used to detect active objects. An object was classified as active if the intersection over union between its bounding box and a hand-object interaction bounding box exceeded 0.8. This method allowed for differentiation between objects that were actively manipulated by the participant and those that were passive within the environment.

We previously evaluated the performance of two object detection models—Detic [25] and UniDet [27]—on our dataset based on their capability to detect a wide variety of object classes. We used a stratified sampling approach, randomly selecting two videos from each participant for each ADL category for annotation. The resulting subset used for evaluating object detection comprised 1482 images, containing 4757 object bounding boxes, which represented approximately 5% of the total dataset [28]. We chose Detic for this study due to its superior performance, achieving a mean Average Precision (mAP) of 0.19 for all objects and 0.30 for active objects [28]. However, our ground truth labeling approach, which involved marking bounding boxes only on objects relevant to the ADLs rather than all objects, led to a high rate of false positives, suggesting that the actual performance of the model might be even better.

The 100DOH model [26] for hand-object interaction was evaluated on 632,180 manually annotated frames from 13 participants and achieved a median F1-score of 0.80 (0.67-0.85), indicating good performance on our dataset [4].

Two approaches were explored for generating a "bag of objects" (BoO) vector: counts and binary presence. The counts method tallied total object occurrences across frames, while binary presence flagged object presence in each frame and took the sum across frames within a snippet. We implemented row-wise min-max scaling to prevent the model from learning specific counts, but rather the proportion of objects instead.

D. ADL Classification and Evaluation

Classifiers including logistic regression (LR), random forest (RF), gradient boosting (GB), extreme gradient boosting

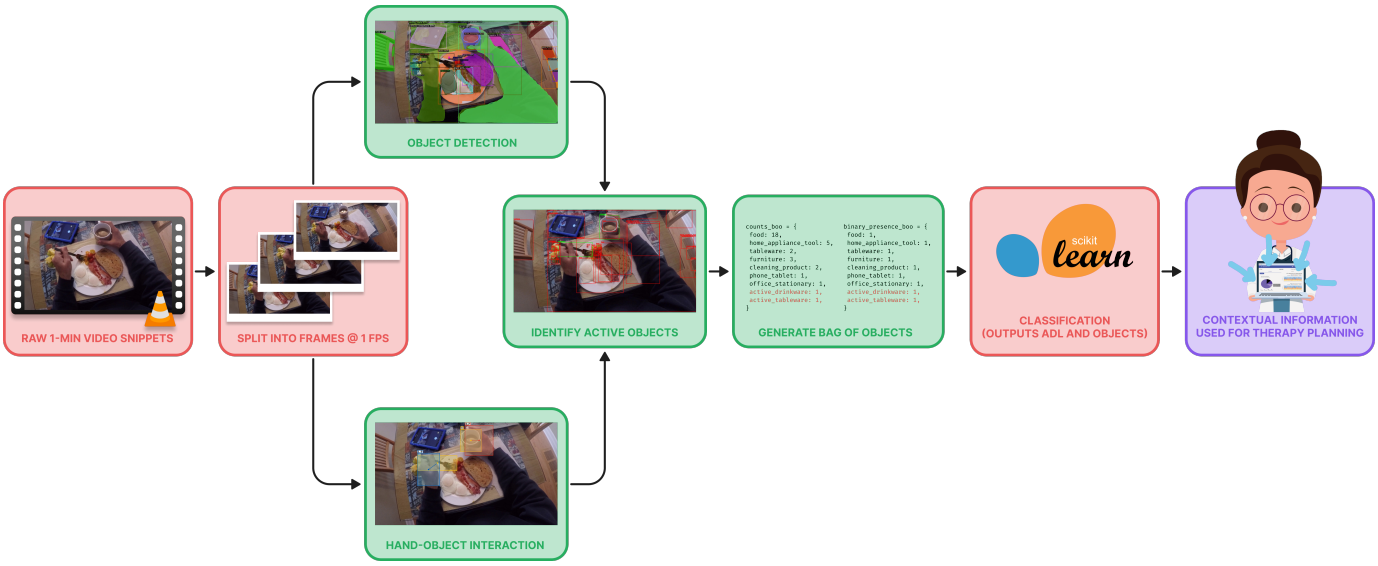

Fig. 1. Process pipeline for detecting ADLs from egocentric videos. Steps in red indicate operations on the entire video snippet while steps in green indicate operations on individual frames. Contextual information from the detected ADLs and objects can be used by clinicians to influence therapy planning.

(XGB), and a multi-layer perceptron (MLP) were trained to classify ADLs. Leave-one-subject-out cross-validation was employed as our evaluation strategy, and our evaluation metrics included mean weighted F1-score by support (this alters the macro-average to account for label imbalance) and the percentage of participants with a weighted F1-score greater than 0.5.

III. Results

The optimization process for our object detection models, while not statistically significant, revealed consistent performance improvements when active objects were considered, though the extent of improvement varied based on the configuration of the BoO vector. We observed an average increase in mean weighted F1-scores of 0.042 ± 0.016 ($p = 0.74$), 0.052 ± 0.016 ($p = 0.68$), and 0.026 ± 0.017 ($p = 0.82$) across models with the inclusion of active objects for counts, binary presence, and counts + binary presence configurations, respectively (Table I).

In a similar vein, we observed a non-significant average increase in mean weighted F1-score of 0.016 ± 0.018 ($p = 0.90$) without active objects and 0.026 ± 0.029 ($p = 0.83$) with active objects when using binary presence over counts BoO vector configuration. Interestingly, we found that the concatenation of counts and binary presence, did not improve performance over just using binary presence. Overall, our best performing model was a LR using a binary presence BoO representation and including active objects.

We also evaluated the sensitivity of our models to object detection performance by training models on ground truth bounding boxes and comparing their performance to object detection predictions (Table II). Model performance was comparable, with XBG, MLP, and GB being the only models that performed better with the ground truth data. This indicates that ADL classifier may be robust to object detection performance.

Overall, LR had the highest performance across all combinations in Table I. Looking at the confusion matrix (Fig. 2) from the best performing model, we observe strong performance on ADLs like "Communication Management", "Meal Preparation and Cleanup", "Self Feeding", "Home Management", and "Functional Mobility". However, the model is less effective at identifying "Grooming & Health Management" and "Leisure & Other Activities", which happen to have the lowest number of samples in the training data.

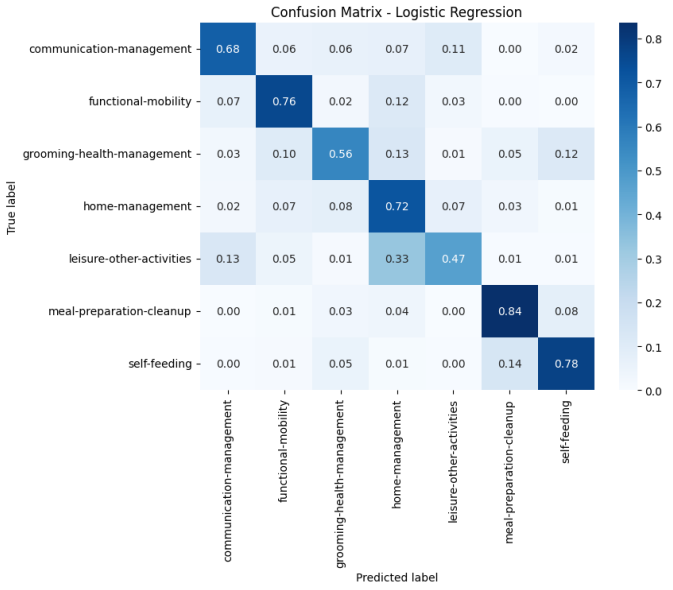

Fig. 2. Confusion matrix for Logistic Regression using binary presence bag-of-objects and including flags for active objects.

Qualitative analysis reveals several cases where predicted labels are incorrect, but may be interpreted as correct based on the label definition. For example, in Fig. 3 (Left) the participant is swallowing pills. This is categorized as "Grooming &

TABLE I
PERFORMANCE OF CLASSIFIERS ACROSS CONFIGURATION SETTINGS.

Configuration	Active Objects	Mean Weighted F1-score	Percentage of Participants >0.5 F1-score
Counts		GB: 0.66 ± 0.23 LR: 0.70 ± 0.14 MLP: 0.65 ± 0.20 RF: 0.65 ± 0.22 XGB: 0.67 ± 0.22	GB: 81% LR: 88% MLP: 81% RF: 75% XGB: 81%
Binary Presence		GB: 0.70 ± 0.18 LR: 0.73 ± 0.15 MLP: 0.65 ± 0.24 RF: 0.64 ± 0.24 XGB: 0.69 ± 0.18	GB: 81% LR: 94% MLP: 81% RF: 81% XGB: 88%
Counts + Binary Presence		GB: 0.69 ± 0.20 LR: 0.72 ± 0.13 MLP: 0.68 ± 0.18 RF: 0.69 ± 0.23 XGB: 0.68 ± 0.17	GB: 81% LR: 94% MLP: 88% RF: 81% XGB: 81%
Counts	✓	GB: 0.69 ± 0.20 LR: 0.73 ± 0.13 MLP: 0.70 ± 0.18 RF: 0.72 ± 0.22 XGB: 0.70 ± 0.20	GB: 88% LR: 94% MLP: 81% RF: 81% XGB: 88%
Binary Presence	✓	GB: 0.73 ± 0.18 **LR: 0.78 ± 0.12** MLP: 0.73 ± 0.22 RF: 0.69 ± 0.25 XGB: 0.74 ± 0.17	GB: 88% **LR: 100%** MLP: 81% RF: 81% XGB: 88%
Counts + Binary Presence	✓	GB: 0.71 ± 0.19 LR: 0.77 ± 0.13 MLP: 0.72 ± 0.16 RF: 0.69 ± 0.23 XGB: 0.70 ± 0.17	GB: 81% LR: 94% MLP: 94% RF: 81% XGB: 88%

Health Management" since medication routines fall under this ADL. However, swallowing pills may also be interpreted as "Self-Feeding" especially since it is being done in the kitchen with drinkware and tableware in the surrounding area.

IV. DISCUSSION

Recent advancements in egocentric activity recognition have leveraged vision transformers and neural networks that integrate multimodal data, including audio [29] and additional sensors [30], to better capture human actions. Advanced domain adaptation techniques are also gaining traction, aiming to improve model generalization across diverse egocentric video datasets without extensive retraining [31]. While these cutting-edge methods may achieve higher overall accuracy, our object-centric approach is designed with practicality in mind, particularly for neurorehabilitation settings. Our methodology employs independently pre-trained object detection and hand-object interaction modules, which not only ensures that our system can easily integrate advancements by swapping out older models for more sophisticated ones as they become available but also maintains focus on object interactions. This focus is particularly beneficial in clinical applications where it's crucial to accommodate the varied compensatory movements that individuals with stroke or SCI may use when performing ADLs. By emphasizing adaptability and scalability, our approach meets the real-world needs of patients and clinicians, ensuring that the technology can be effectively used in diverse and dynamic real-life environments.

TABLE II
PERFORMANCE OF CLASSIFIERS TRAINED ON BOO VECTORS GENERATED USING BINARY PRESENCE WITH ACTIVE OBJECTS INCLUDED FROM GROUND TRUTH BOUNDING BOXES VERSUS DETECTED OBJECTS ON A SUBSET (5%) OF THE DATA.

Training Data	Mean Weighted F1-score	Percentage of Participants >0.5 F1-score
Ground Truth	GB: 0.58 ± 0.16 LR: 0.62 ± 0.17 MLP: 0.41 ± 0.23 RF: 0.64 ± 0.21 XGB: 0.61 ± 0.23	GB: 62% LR: 69% MLP: 31% RF: 61% XGB: 69%
Detected Objects	GB: 0.54 ± 0.20 LR: 0.65 ± 0.16 MLP: 0.35 ± 0.23 RF: 0.66 ± 0.22 XGB: 0.56 ± 0.21	GB: 56% LR: 75% MLP: 19% RF: 75% XGB: 59%

The inclusion of objects, especially active objects, can accurately detect ADLs from egocentric video footage of individuals with stroke and SCI in home settings. Our best model, LR using binary presence and active objects, achieved a mean weighted F1-score of 0.78 ± 0.12 using leave-one-subject-out cross-validation, with 100% of participants scoring above 0.5.

We found that in general, models were more performant with the binary presence BoO vector representation as opposed to object counts. This may be due to inaccuracies in object detection predictions, which may have added additional noise to the counts models, causing them to overfit to the counts of particular objects. As a result, models trained using binary presence may be robust to inaccurate object detections at the frame level since the correct objects are detected at some point in the frames of a video snippet.

Our comparative analysis using object detection predictions versus ground truth data highlighted that certain classifiers, specifically RF and LR, performed better with detection data than ground truth. This suggests that our ADL classifiers are robust to variations in object detection performance, pointing to the potential for practical deployment even with imperfect detection systems. However, the differential performance between the two data representations indicates room for exploring more sophisticated data representations that could yield better performance.

Our best performing model showed limitations in accurately classifying "Grooming & Health Management" and "Leisure & Other Activities". These classes often have the most considerable variation of activities since "Leisure & Other Activities" can encompass any activity that does not fit into the other classes, and "Grooming & Health Management" can include activities like brushing teeth and exercising simultaneously. Therefore, the poor performance can likely be attributed to data imbalances, with these two ADLs having the smallest number of samples in the dataset (172 and 165 instances, respectively). Future work should focus on refining ADL classes to enhance discrimination accuracy, perhaps by redefin-

Prediction:
Self-Feeding

Ground Truth:
Grooming & Health Management

Detected Objects:
'active_drinkware': 0.1538
'active_phone_tablet': 0.0769
'active_electronics': 0.0769
'phone_tablet': 0.1538
'other': 0.2307
'office_stationary': 0.0769
'furniture': 0.0769
'furnishing': 0.3076
'drinkware': 0.6923
'home_appliance_tool': 0.5384
'cleaning_product': 1.0,
'bag': 0.0769
'electronics': 0.0769
'kitchen_appliance': 0.1538
'house_fixtures': 0.3846
'tableware': 1.0

Prediction:
Home Management

Ground Truth:
Grooming & Health Management

Detected Objects:
'clothing_accessory': 0.0769
'phone_tablet': 0.1538
'other': 0.9230
'office_stationary': 0.0769
'footwear': 0.6923
'furniture': 0.0769
'furnishing': 0.0769
'drinkware': 0.0769
'home_appliance_tool': 1.0
'food': 0.0769
'clothing': 0.3076
'cleaning_product': 0.0769
'electronics': 0.0769
'wheelchair_walker': 0.1538
'sports_equipment': 0.6153
'tv_computer': 0.0769
'house_fixtures': 0.2307
'tableware': 0.2307

Prediction:
Meal Preparation & Cleanup

Ground Truth:
Meal Preparation & Cleanup

Detected Objects:
'active_food': 0.7692
'clothing_accessory': 0.0769
'other': 0.1538
'furnishing': 0.0769
'drinkware': 0.3846
'home_appliance_tool': 0.5384
'food': 1.0
'cleaning_product': 1.0
'toiletries': 0.0769
'kitchen_utensils': 1.0
'sink': 1.0
'house_fixtures': 0.8461
'tableware': 1.0
'bathroom_fixture': 0.6923

Fig. 3. Qualitative evaluation of ADL calssification displaying sample frames from the video snippets, followed by the ADL prediction, ADL ground truth, and the min-max normalization of detected objects from our best performing model, LR with active objects and binary presence.

ing categories or introducing more granular classes related to specific object interactions.

Analysis of the confusion matrix indicates inadequate performance for "Grooming & Health Management" and "Leisure & Other Activities". More granular ADL classes that have objects that belong to the same functional categories (e.g., "Meal Preparation and Cleanup" has many instances of "Kitchen Utensils") should be considered to make them easier to discriminate.

Naturally, this means that a significant limitation in this study is the balance of classes in the dataset. However, we have tried to account for this by choosing evaluation metrics that lend themselves well to imbalanced data. While objects can provide insights into the activity being performed, they may also lead to misinterpretation in situations where activities occur in unexpected contexts. For instance, a participant exercising while seated at the dining table, as depicted in Fig. 3 (Middle), illustrates this potential for misalignment between observed objects and the actual activity taking place. While we accounted for imbalances in the dataset using cost-sensitive learning, other strategies for handling data imbalances could be considered, such as oversampling the under-represented classes with techniques like SMOTE [32], or considering additional dimensions of class imbalance [33]. The extraction of additional information can also be explored to gather more context on how the objects are being used using vision

language models to more accurately predict the ADL being performed [34].

V. Conclusion

Using an object-centric BoO approach, we demonstrate the use of object detection as a proxy for accurately identifying ADLs in egocentric video snippets of 1-min in length. Our findings reveal that the binary presence BoO vector representation consistently outperforms object counts, likely due to its robustness against inaccuracies, and false positives in particular, that may be present in object detection predictions. LR with binary presence and active objects achieved the highest performance with a mean weighted F1-score of 0.78 $\pm$ 0.12 using a leave-one-subject-out cross validation. Using this method, we are able to provide clinicians with accurate information about ADLs and object interactions "in the wild".

Despite our acceptable performance on 7 different ADL classes, challenges persist, particularly in dealing with class imbalances that affect the model's ability to accurately classify less frequent ADLs such as "Grooming & Health Management" and "Leisure & Other Activities". However, this may be due to the range of activities and objects that could comprise these ADLs, outlining a key limitation with this particular methodology.

Future work should focus on refining the categorization of detected objects using automated methods such as clustering or semantic categorization in the consolidation phase to reduce manual intervention and potentially increase the scalability of our system. Additionally, further research into different feature engineering techniques besides counts or binary presence could provide even greater accuracy and efficiency in real-world applications. Overall, our work demonstrates the potential for providing clinicians with actionable contextual information to guide therapy planning using information about their patients' actual hand use at home.

References

[1] K. D. Anderson, "Targeting recovery: priorities of the spinal cord-injured population," *J. Neurotrauma*, vol. 21, no. 10, pp. 1371–1383, Oct. 2004.

[2] J. Purton, J. Sim, and S. M. Hunter, "Stroke survivors' views on their priorities for upper-limb recovery and the availability of therapy services after stroke: a longitudinal, phenomenological study," *Disabil. Rehabil.*, vol. 45, no. 19, pp. 3059–3069, Sep. 2023.

[3] 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.

[4] 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*, vol. 39, no. 23-24, pp. 1697–1707, Dec. 2022.

[5] A. Kadambi, A. Bandini, R. D. Ramkalawan, S. L. Hitzig, and J. Zariffa, "Designing an egocentric video-based dashboard to report hand performance measures for outpatient rehabilitation of cervical spinal cord injury," *Top. Spinal Cord Inj. Rehabil.*, vol. 29, no. suppl, pp. 75–87, Nov. 2023.

[6] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and Others, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.

[7] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The EPIC-KITCHENS dataset: Collection, challenges and baselines," *arXiv [cs.CV]*, Apr. 2020.

[8] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4D: Around the world in 3,000 hours of egocentric video," *arXiv [cs.CV]*, Oct. 2021.

[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 20–36.

[10] K. Tsmzhenyangli, E. Gavrilyuk, M. Gavves, and G. M. Cees, "VideoLSTM convolves, attends f lows action recognition," *Computer Vision Image Understanding*, vol. 166, pp. 41–50, 2018.

[11] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," *arXiv [cs.CV]*, Nov. 2018.

[12] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," *arXiv [cs.CV]*, Nov. 2017.

[13] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *arXiv [cs.CV]*, May 2017.

[14] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," *arXiv [cs.CV]*, Dec. 2018.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv [cs.CL]*, Jun. 2017.

[16] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," *arXiv [cs.CV]*, Jun. 2021.

[17] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" *arXiv [cs.CV]*, Feb. 2021.

[18] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 407–414.

[19] R. Granada, J. Monteiro, N. Gavenski, and F. Meneguzzi, "Object-based goal recognition using real-world data," in *Advances in Soft Computing*. Springer International Publishing, 2020, pp. 325–337.

[20] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2012.

[21] R. Nabiei, M. Parekh, E. Jean-Baptiste, P. Jancovic, and M. Russell, "Object-centred recognition of human activity," in *2015 International Conference on Healthcare Informatics*. IEEE, Oct. 2015, pp. 63–68.

[22] L. Lonini, A. Gupta, K. Kording, and A. Jayaraman, "Activity recognition in patients with lower limb impairments: Do we need training data from each patient?" in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Aug. 2016, pp. 3265–3268.

[23] 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.

[24] Aota, *Occupational Therapy Practice Framework: Domain & Process*. American Occupational Therapy Association, Incorporated, Sep. 2020.

[25] 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.

[26] 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.

[27] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," *arXiv [cs.CV]*, Feb. 2021.

[28] A. Kadambi and J. Zariffa, "Providing hand use context for outpatient neurorehabilitation with egocentric object detection," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, Oct. 2023, pp. 1–4.

[29] A. Papadakis and E. Spyrou, "A multi-modal egocentric activity recognition approach towards video domain generalization," *Sensors (Basel)*, vol. 24, Apr. 2024.

[30] Y. Hao, A. Kanezaki, I. Sato, R. Kawakami, and K. Shinoda, "Egocentric human activities recognition with multimodal interaction sensing," *IEEE Sens. J.*, vol. 24, no. 5, pp. 7085–7096, Mar. 2024.
[31] X. Liu, T. Lei, and P. Jiang, "Fine-grained egocentric action recognition with multi-modal unsupervised domain adaptation," in *2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, vol. 6. IEEE, Feb. 2023, pp. 84–90.
[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *JAIR*, vol. 16, pp. 321–357, Jun. 2002.
[33] O. Wu, "Rethinking class imbalance in machine learning," *arXiv [cs.LG]*, May 2023.
[34] G. Dai, X. Shu, and W. Wu, "GPT4Ego: Unleashing the potential of pre-trained models for zero-shot egocentric action recognition," *arXiv [cs.CV]*, Jan. 2024.