Generalizability of Hand-Object Interaction Detection in Egocentric Video across Populations with Hand Impairment

Meng-Fen Tsai, Rosalie H. Wang and José Zariffa, *Senior Member, IEEE*

***Abstract*— Stroke survivors often experience unilateral sensorimotor impairment. The restoration of upper limb function is an important determinant of quality of life after stroke. Wearable technologies that can measure hand function at home are needed to assess the impact of new interventions. Egocentric cameras combined with computer vision algorithms have been proposed as a means to capture hand use in unconstrained environments, and have shown promising results in this application for individuals with cervical spinal cord injury (cSCI). The objective of this study was to examine the generalizability of this approach to individuals who have experienced a stroke. An egocentric camera was used to capture the hand use (hand-object interactions) of 6 stroke survivors performing daily tasks in a home simulation laboratory. The interaction detection classifier previously trained on 9 individuals with cSCI was applied to detect hand use in the stroke survivors. The processing pipeline consisted of hand detection, hand segmentation, feature extraction, and interaction detection. The resulting average F1 scores for affected and unaffected hands were 0.66 ± 0.25 and 0.80 ± 0.15, respectively, indicating that the approach is feasible and has the potential to generalize to stroke survivors. Using stroke-specific training data may further increase the accuracy obtained for the affected hand.**

I. Introduction

Individuals who have had a stroke experience hemiplegia or hemiparesis, which is a unilateral motor deficit on the contralateral side of the brain lesion. One of the determinants of quality of life and independence after stroke is upper limb function [1]. An estimated 65% of stroke survivors experience difficulties in their activities of daily living (ADLs) as a result of upper limb impairment, despite medication and rehabilitation [2-4]. Capturing the upper limb function of stroke survivors in their daily life is vital to quantifying the impact of new interventions and to designing personalized rehabilitation plans.

In order to measure the upper limb function of stroke survivors in their living environment, accelerometers have been used to capture upper limb movements [5]. However, accelerometers do not document whether the detected upper limb movement belongs to a functional task [6, 7]. In addition, most of the studies using accelerometers for detecting upper limb movements place the devices on the wrists, limiting their ability to capture the details of hand movements. Furthermore, studies that recorded hand movements with accelerometers worn on fingers were carried out with well-defined tasks in laboratories rather in home or community environments [8, 9]. There is still a need for wearable technologies that can measure how individuals with stroke use their hands at home, in order to better reflect the impact of interventions on daily life.

Videos from wearable cameras (egocentric video) have recently been proposed as a means to capture hand use in unconstrained environments. Building on prior work in computer vision approaches to recognize objects [10-12] and activities [13] in egocentric video, a machine learning-based system was used to detect interactions between the hands and objects in the environment [14, 15]. Hand-object interactions were defined as participants manipulating an object with their hand for a functional purpose. This definition was used as the basis for a frame-by-frame binary classification task (interaction / no interaction). It was postulated that detecting hand-object interactions could serve as the basis for measures of recovery of hand use, for example by quantifying the instances of functional uses of the impaired hand over time. In this previous study, Likitlersuang et al. [14] showed promising results for detecting the hand use of individuals with cervical spinal cord injury (cSCI) using this approach.

In order to extend this novel system to a wider population of individuals with hand impairments, the algorithms previously trained on individuals with cSCI were evaluated in the present study for their ability to detect hand-object interactions in stroke survivors. In particular, we sought to determine the generalizability of the trained models across populations.

II. Methods

Egocentric cameras (GoPro Hero 4 and 5, GoPro Inc., CA, USA), which record video from a first-person angle, were used to record 38 daily tasks in a home simulation laboratory. The videos were recorded at 1280x720 resolution and at 30 frames per second. However, they were analyzed in a reduced resolution of 640x360 pixels.

In this study, two groups of participants with hand impairment were involved: individuals with cSCI and stroke survivors. The inclusion criteria for study participants in both groups were as follows.

For the individuals with cSCI, the inclusion criteria were: 1) a neurological level of injury between C2-T1; 2) an impairment grade between A and D in the American Spinal Injury Association Impairment Scale (AIS); 3) no wrist or

*Research supported by the Heart and Stroke Foundation of Canada (G-18-0020952).

Meng-Fen Tsai (e-mail: mf.tsai@mail.utoronto.ca), Rosalie Wang (email: rosalie.wang@utoronto.ca) and José Zariffa (e-mail: jose.zariffa@utoronto.ca) are with KITE - Toronto Rehabilitation Institute – University Health Network, Toronto, Ontario, Canada. Meng-Fen Tsai and José Zariffa are also with the Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada. Rosalie Wang is with the Department of Occupational Science and Occupational Therapy, University of Toronto.

hand deformities/injuries; 4) impaired but not completely absent hand function. No restrictions were placed on injury type (traumatic or non-traumatic) or duration after injury. The cSCI data used here is part of the dataset previously reported in [14].

For the stroke survivors, the inclusion criteria were: 1) at least six months post-stroke; 2) self-reported impact of affected hand on ADLs; 3) impaired but not absent hand function; 4) no subluxation or significant pain when using upper limb; 5) no other neuromusculoskeletal disease affecting upper limb movement other than stroke.

The study was approved by the Research Ethics Board of the University Health Network. Signed consent was obtained from each participant before enrolling them in this study.

A. Dataset

The training dataset consisted of 82,331 video frames collected in a previous study, according to procedures described in detail in [14]. The dataset included frames of interaction (57%) and frames of no interaction (43%) from videos of 9 individuals with cSCI executing ADLs in the simulated home environment.

The testing dataset consisted of 27,066 frames in total from videos of 6 individuals with stroke executing ADLs in the home simulation laboratory. Both the training and testing sets included videos across 6 rooms with different settings (e.g. living room, kitchen and bedroom) and 38 tasks. The participants were instructed to perform the tasks in the manner they usually did at home. The testing dataset consisted of 70% of interactions and 30% of non-interactions.

B. Hand Detection

Before detecting interactions, the hands must be located in the image. A convolution neural network (CNN) - You Only Look Once version 2 (YOLOv2) - was previously retrained to detect hands using data from individuals with cSCI [16], and was applied in this study. YOLOv2 generated the hand coordinates (bounding boxes) in each image (Fig. 1). If the intersection over union (IoU) of the generated bounding box and manually labeled one was above 0.5, the result was considered a true positive. Otherwise, it was categorized as a false positive.

While YOLOv2 was used to detect hands of stroke survivors in this study due to high accuracy and improved speed, in the previous study, a Faster Region-CNN was used to find the hand bounding boxes, as described in [14]. Both hand detection methods provided the hand coordinates and identified right and left hands of the user.

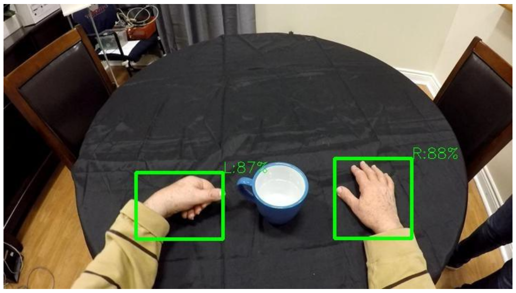

Figure 1. Hand detection example: the detected bounding boxes of two hands. Numbers denote confidence levels of the detections.

C. Hand Segmentation

Some of the features used for the interaction detection step (see section II. D) required the hand to be segmented. Two features were used to segment a hand – skin colour and edges. A generic mixture-of-Gaussians skin colour model [17] and a Structured Forest edge detection approach [18] were used to find hand pixels and to delineate edges within the bounding boxes of the hands (Fig. 2). A hand was segmented by combining the given hand pixels and edges within the bounding boxes, as described in [14].

D. Hand-Object Interaction Detection

Three types of features were extracted for detecting hand-object interactions: hand shape, object colour and object motion. The object colour and motion features were generated by comparing the regions of the segmented hand, near the hand (within the bounding box) and background (outside of the bounding box) [14].

The hand shape was extracted within the bounding box of a hand using the Histogram of Oriented Gradients (HOG) [19, 20]. The bounding boxes were re-sized to 10% of the width and the length of the original image. The HOG were computed in the re-sized bounding boxes using 16x16 pixels as a cell, and 2x2 cells per block. Principal component analysis (PCA) was applied to reduce the dimensionality of the HOG feature vector, with the first 60 principal components retained.

The colour features were generated using the Hue, Saturation and Value (HSV) colour space. The HSV colour histograms were compared between the hand (region (a) in Fig. 3) and the region near the hand in the bounding box (region (b) in Fig. 3), and between the region near the hand and the background (region (c) in Fig. 3) , using the Bhattacharyya distance. These two differences were used as colour features, based on an idea that the presence of an object near the hand might result in colour differences with a hand or background.

The motion feature was generated based on optical flow. The rationale was that the movement direction and velocity of manipulated objects would be similar to that of the hand. Thus, the motion feature consisted of two arrays. One is the subtraction of the optical flow histogram of the region around the hand from that of the hand, and the other is the subtraction of the histogram of the region around the hand from that of the background.

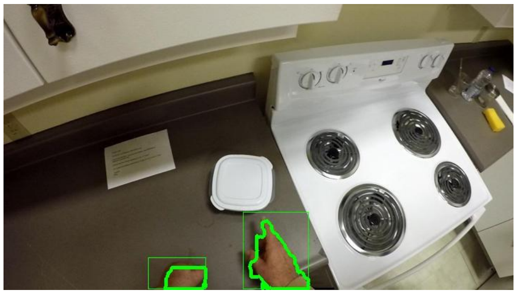

Figure 2. Hand segmentation example. The bounding boxes and segmentation regions are shown.

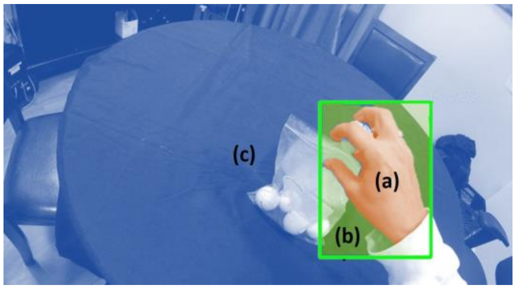

Figure 3.The three regions used in the colour and motion features: the hand (a), the region around hand in the detected bounding box (b) and the background (c).

The feature vectors were fed into a random forest classifier with 150 trees to detect functional interactions between the hands and objects (binary classification). The classifier was trained on the 9 individuals with cSCI and tested on the 6 participants after stroke.

Lastly, a filtering step was implemented to smooth out results across consecutive frames. The predictions of the classifier and the manual annotations were passed through a moving average filter with a window of 120 frames. Next, the filtered binary outputs were normalized by subtracting the minimum value over the task and then dividing by the range of values observed in the task. The threshold for an interaction was set at 0.5. After filtering and normalizing, the F1 score was used to evaluate how well the algorithm could detect hand-object interactions.

III. Results

There were 8 male and 1 female participants with cSCI, and 4 male and 2 female participants after stroke involved in this study. The average age for the cSCI group and stroke group were 52.1 ± 13.1 and 56.8 ± 19.3, respectively. The severity level of upper limb impairment in the two groups were reported using the Upper Extremity Motor Score (UEMS) and Fugl-Meyer Assessment for Upper Extremity (FMA-UE) for participants with cSCI and stroke, respectively. The median (interquartile range) UEMS was 18 (14-20) and the median (interquartile range) FMA-UE was 37 (27-55). The participants with cSCI ranged from AIS A to D. As for the participants with stroke, the participants included individuals with severe, moderate and mild upper limb impairment.

A. Hand Detection

The 27,066 frames from 6 stroke survivors were labeled and tested. The labeled frames consisted of 16,821 frames with affected hands and 21,549 frames with unaffected ones. The average IoU was 0.54 ± 0.07 and 0.68 ± 0.06 for affected and unaffected hands, respectively. The average F1 score for affected hands and unaffected hands were 0.77 ± 0.24 and 0.82 ± 0.17, respectively. The average precision and recall of the affected hand was 0.57 ± 0.36 and 1.00 ± 0.00. For the unaffected hands, the average precision and recall was 0.73 ± 0.21 and 1.00 ± 0.00, respectively.

B. Hand-Object Interaction

Each participant with stroke had 6 to 7 tasks across 6 rooms reported in this study. The F1 scores, precisions and recalls for each subject and average results are shown in Table I. The average F1 score of hand-object interaction for affected and unaffected hands were 0.66 ± 0.25 and 0.80 ± 0.15, respectively. The mean and standard deviation of F1 score, precision and recall of unaffected hands were all higher than the affected ones. According to the participants' FMA-UE score, most of the participants with low F1 scores had moderate to severe upper limb impairments.

TABLE I. Hand-object interaction

Subject ID	FMA-UE score *	Affected hand			Unaffected hand		
		F1 score	***P***	***R***	***F1 score***	***P***	***R***
1	27	0.67	0.45	0.80	0.83	0.76	0.81
2	27	0.18	0.05	0.68	0.86	0.85	0.91
3	56	0.72	0.44	0.89	0.77	0.59	0.92
4	24	0.68	0.36	0.99	0.83	0.94	0.84
5	47	0.81	0.82	0.87	0.99	0.99	0.99
6	66	0.88	0.93	0.86	0.53	0.27	0.97
Mean ± SD		0.66 ± 0.25	0.51 ± 0.32	0.85 ± 0.10	0.80 ± 0.15	0.73 ± 0.27	0.91 ± 0.07

*P: Precision. R: Recall. *: Assessed on the affected side.*

IV. Discussion

In our previous study [14], a similar approach was applied to 9 individuals with cSCI using a leave-one-subject-out cross-validation process, and the average F1 score for right and left hand were 0.73 ± 0.15 and 0.74 ± 0.15. In this study, the model trained on cSCI was applied to 6 stroke survivors and the F1 scores for the unaffected hands of the stroke survivors were maintained. For the F1 scores of affected hands, the results were variable. Participant 2, who had the lowest interaction detection F1 score, used their affected hand in only one task while the rest were carried out by the unaffected hand. The number of frames containing the affected hand was small. Analyzing more frames might be required for more reliable estimates. Furthermore, the resting hand shape for this participant was a slightly closed hand, which might be classified into an interacting hand and lead to a high incidence of interaction false positives. The results demonstrate that the wearable system has the potential to detect hand-object interactions across two populations with unilateral and bilateral hand impairments. However, with stroke survivors with moderate to severe upper limb function, the results can be improved. To overcome the small number of affected hand images, recording tasks that require bilateral interactions may be helpful to increase the diversity of the training sample.

There were two limitations in this study. First, the number of frames involving affected hands was smaller than the number of frames with unaffected ones. The difference in the number of frames between the two hands might reflect the nonuse phenomenon. Stroke survivors tend to use their unaffected hands, which have better fine motor skills, to manipulate objects in ADLs. Furthermore, individuals with severe upper limb function impairment may not be able to lift their hands to a table. Affected hands being occluded under a table or not visible in the frame is a factor that could reduce the performance of hand-object interaction detection. Conversely, individuals with cSCI typically have two affected hands and may not show a preference of using only one hand. Thus, in the cSCI study [14], the average F1 scores for the two hands were similar on average (however, some individuals did display an asymmetry in the interaction detection performance). Second, the testing (stroke) dataset consisted only of objects manipulation tasks and contained limited data without any hand use. Adding data without hand use would overcome the bias of the testing dataset.

A further revision is necessary to increase the accuracy of detecting hand-object interactions in the affected hands of stroke survivors. The stroke survivors with severe hand impairment usually keep their affected hand fully flexed in a fist. In contrast, individuals with cSCI may not have their hand fully flexed, depending on the injury characteristics and contractures. The different hand shapes during object manipulations between stroke survivors and individuals with cSCI might influence the detection accuracy when examining the affected hands of stroke survivors.

Although the average F1 score of unaffected hands was consistent with the previous cSCI results, the results for the affected hands were lower in this pilot sample. Further data collection will be needed before conducting statistical comparisons to confirm this trend. The algorithm might perform better on affected hands if the classifiers were trained on stroke data, in order to adapt to the different hand shapes prevalent in different conditions.

While the use of egocentric video in rehabilitation applications raises some privacy-related challenges, previous studies have found these to be acceptable to individuals with cSCI. [21]. Future work will be required to confirm these findings with individuals who have experienced a stroke.

V. Conclusion

Detecting hand-object interactions in one population with hand impairment using an algorithm trained on another population (stroke and sSCI, respectively) was feasible in most participants. The F1 score, precision and recall of the hand-object interaction detection in the affected hand of stroke survivors were all lower than in the unaffected ones. Transfer learning approaches and population-specific training would be beneficial to explore for better detection in the population with unilateral hand impairment.

Acknowledgment

The authors would like to thank all of the study participants, as well as Ryan J. Visée and Jirapat Likitlersuang, whose work paved the way of this study.

References

[1] Nichols-Larsen, D.S., et al., Factors influencing stroke survivors' quality of life during subacute recovery. Stroke, 2005. **36**(7): p. 1480-1484.

[2] Lum, P.S., et al., Gains in upper extremity function after stroke via recovery or compensation: potential differential effects on amount of real-world limb use. Topics in stroke rehabilitation, 2009. **16**(4): p. 237-253.

[3] Mayo, N.E., et al., Activity, participation, and quality of life 6 months poststroke. Archives of physical medicine and rehabilitation, 2002. **83**(8): p. 1035-1042.

[4] Dobkin, B.H., Rehabilitation after stroke. New England Journal of Medicine, 2005. **352**(16): p. 1677-1684.

[5] Urbin, M., K.J. Waddell, and C.E. Lang, Acceleration metrics are responsive to change in upper extremity function of stroke survivors. Archives of physical medicine and rehabilitation, 2015. **96**(5): p. 854-861.

[6] Hayward, K.S., et al., Exploring the role of accelerometers in the measurement of real world upper-limb use after stroke. Brain Impairment, 2016. **17**(1): p. 16-33.

[7] van der Pas, S.C., et al., Assessment of arm activity using triaxial accelerometry in patients with a stroke. Archives of physical medicine and rehabilitation, 2011. **92**(9): p. 1437-1442.

[8] Hester, T., et al. Using wearable sensors to measure motor abilities following stroke. in International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06). 2006. Cambridge, MA, USA: IEEE.

[9] Patel, S., et al., A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology. Proceedings of the IEEE, 2010. **98**(3): p. 450-461.

[10] Fathi, A., X. Ren, and J.M. Rehg. Learning to recognize objects in egocentric activities. in Computer Vision and Pattern Recognition. 2011. Colorado Springs, CO, USA: IEEE.

[11] Lee, Y.J., J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. in IEEE conference on computer vision and pattern recognition. 2012. Providence, RI, USA: IEEE.

[12] Ren, X. and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. in IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010. San Francisco, CA, USA: IEEE.

[13] Ryoo, M.S. and L. Matthies. First-person activity recognition: What are they doing to me? in IEEE conference on computer vision and pattern recognition. 2013. Portland, OR, USA.

[14] Likitlersuang, J., et al., Egocentric video: a new tool for capturing hand use of individuals with spinal cord injury at home. Journal of neuroengineering and rehabilitation, 2019. **16**(1): p. 83.

[15] Likitlersuang, J. and J. Zariffa, Interaction Detection in Egocentric Video: Toward a Novel Outcome Measure for Upper Extremity Function. IEEE journal of biomedical and health informatics, 2018. **22**(2): p. 561-569.

[16] Visée, R.J., J. Likitlersuang, and J. Zariffa, An Effective and Efficient Method for Detecting Hands in Egocentric Videos for Rehabilitation Applications. arXiv preprint arXiv:1908.10406, 2019.

[17] Jones, M.J. and J.M. Rehg, Statistical color models with application to skin detection. International Journal of Computer Vision, 2002. **46**(1): p. 81-96.

[18] Dollár, P. and C.L. Zitnick, Fast edge detection using structured forests. IEEE transactions on pattern analysis machine intelligence, 2015. **37**(8): p. 1558-1570.

[19] Cai, M., K.M. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. in 2015 IEEE International Conference on Robotics and Automation (ICRA). 2015. Seattle, WA, USA: IEEE.

[20] Dalal, N. and B. Triggs. Histograms of oriented gradients for human detection. in IEEE computer society conference on computer vision and pattern recognition. 2005. Boston, MA, USA: IEEE.

[21] Likitlersuang, J., et al., Views of individuals with spinal cord injury on the use of wearable cameras to monitor upper limb function in the home and community. Journal of Spinal Cord Medicine, 2017. **40**(6): p. 706-714.